A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

ENHANCING THE PREDICTION OF SHOT SUCCESS IN NBA BASKETBALL GAMES USING MACHINE LEARNING TECHNIQUES – FEEDFORWARD NEURAL NETWORK

SEBASTIAN MANI VARADAPPA

Work project carried out under the supervision of:

Yufei Shen

12/02/2024

Abstract

The advent of data-driven decision-making has sparked a transformation in the sports industry, where the precision of predictive models now serves as a pivotal factor in both team success and financial viability. This thesis examines Machine Learning and Deep Learning models for predicting NBA shot success, with team members developing Random Forest, XGBoost, Feedforward and Recurrent Neural Network models. Notably, the Recurrent Neural Network, previously unapplied in this context, emerged with superior predictive accuracy. This study's primary contribution is unveiling the RNN's potential for shot prediction, paving the way for its future integration into sports strategic planning and business analytics.

Keywords: Predictive Modeling, Machine Learning, Deep Learning, NBA Basketball, shot success, Random Forest, XGBoost, Feedforward Neural Network, LSTM Neural Network

Supported by Nova School of Business and Economics.

Acknowledgements

Special thanks to our advisor, Yufei Shen, whose dedication and insightful advice greatly enhanced this thesis. His enthusiasm and commitment made this journey an enjoyable experience. We also extend our gratitude to Paulo Marques, representing the Nova Data Science Knowledge Center, for providing crucial computational resources.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

GROUP PART

1. Introduction

1.1 Background

The sports industry has undergone a data revolution, with analytics fundamentally altering the landscape of competition, performance optimization, and fan engagement. The adoption of predictive modeling in sports has been transformative, enabling a leap from traditional, intuition-based decision-making to a more empirical, data-driven approach. By harnessing vast datasets that capture the minutiae of player movements, game dynamics, and myriad other variables, teams and organizations can now discern patterns and correlations that drive strategic decisions. The utility of such analytics extends across various sporting disciplines, enhancing training regimens, refining tactical approaches, and even shaping fan experiences through personalized engagement (Rein et al. 2016). This paradigm shift towards data-centric sports management and strategy formulation has not only elevated the level of play but also opened new avenues for revenue generation and business growth within the sports sector.

In the National Basketball Association (NBA), the introduction of Machine Learning (ML) and Deep Learning (DL) techniques has led to notable enhancements in team performance and player capabilities. These models are most often centered around predicting shot success. While various performance indicators contribute to the success of a basketball team - ranging from coaching strategies to player fitness - shooting accuracy stands out as particularly crucial (Zhang et al. 2019; Huyghe et al. 2022).

Notably, the success rate of shots has increased over the last decade along with data-driven strategic shifts impacting the rate of shot attempts from different zones on the court, as well as the types of shots that players are now optimizing for (Wang and Zemel 2017). For example, there has been a

significant rise in the accuracy of three-point shots, coinciding with an increased frequency of attempts from beyond the arc. This evolution in shooting strategies reflects the profound impact that analytics has had on the sport, leading to more calculated and effective scoring tactics.

Crucially, as sport evolves, so too do the strategies derived from data analysis, creating a cycle of adaptation and counter-adaptation. This is illustrated by a recent shift from 3-pointers towards 2-pointers, which is likely correlated with defensive strategies focusing more on preventing 3-point attempts, thereby allowing more 2-pointers (Vicent et al. 2021). Hence, this dynamic interplay turns predictive modeling into a continuous effort, as strategies are consistently refined to maintain a competitive edge.

Analytics, therefore, is not merely a tool for informing current strategies but a means to anticipate and adapt to future shifts in the sport. In this landscape of ever-evolving tactics and data-driven strategies, predictive modeling remains an indispensable component in the quest for competitive advantage, and consequently, a financially sustainable basketball team within the highly competitive NBA league. It is within this context that our thesis aims to explore and contribute to the ongoing effort to refine the predictive models that aim to predict shot success in basketball.

1.2 Problem Statement

The NBA is a multi-billion-dollar industry where the survival and success of teams hinge crucially on well-informed strategic decisions. In the last section we have seen that in recent years, these strategies have become increasingly reliant on data analysis. As a core aspect of this data-driven approach, the ability to accurately predict shot success in games is paramount. Consequently, teams' strategies and potential victories often depend on the extent to which these models accurately predict shot success. The challenge lies in the fact that shot success is influenced by a multitude of

factors, including those that are difficult to quantify, such as the unpredictable actions of players (Meehan 2017).

This paper aims to address the challenge of improving shot success prediction in the NBA by enhancing ML methods. We aim to do so by addressing limitations found in previous studies related to the volume of data points used, the breadth of feature selection, and the architecture of the models. By focusing on refining these models, the research aims to provide NBA teams with more accurate tools for strategy formulation, thereby enhancing their competitive edge in the league. This effort is not just crucial for the teams' tactical advancements but also significant for the broader context of sports analytics, where the integration of advanced predictive techniques continues to redefine the landscape of competitive sports.

1.3 Research Contribution and Business Implications

This research contributes significantly to the field of sports analytics by enhancing the understanding and application of machine learning (ML) and deep learning (DL) in predicting NBA shot success. As discussed in the previous section, our paper address limitations found in previous studies related to the volume of data points used, the breadth of feature selection, and the architecture of the models. Across the literature, studies such as those by Meehan (2017), Oughali et al. (2019), Wright et al. (2016), Harmon et al. (2016), and Shah and Romijnders (2016) have predominantly utilized data from only one NBA season to train their models. This approach potentially limits the algorithms, particularly those employing DL, from identifying all relevant patterns in the data. Furthermore, these studies are often constrained by a narrow selection of features. For example, Meehan (2017) only incorporated numerical features and categorical variables with two possible values, while Oughali et al. (2019) utilized a mere four features in their

models. Given the complexity of factors influencing shot success, a model relying on such a limited number of features is overly simplistic.

Another critical aspect that most studies have overlooked is the optimization of hyperparameters. Many rely on default settings; for instance, Oughali et al. (2019) used the default hyperparameters of their recurrent neural network (RNN). For DL models, which are sensitive to hyperparameter settings, this lack of optimization can prevent the algorithms from reaching their full predictive potential.

To address these limitations, our study uniquely employs data spanning five NBA seasons (2019-2023), diverging from the typical single-season analyses. We also expand on feature selection, utilizing a total of 21 features, both numerical and categorical, to predict shot success. This more comprehensive set of features is expected to provide a deeper insight into the dynamics influencing shot outcomes.

Furthermore, our approach includes extensive hyperparameter optimization during the model building process. We consider a wide range of possible values and test combinations at a more granular level than in previous studies. This thorough optimization process is aimed at fully exploiting the potential of our ML models and enhancing the accuracy of shot success predictions in NBA games.

Aside from its academic value in advancing basketball analytics, this thesis also has significant practical implications across various sectors within the sports industry. Firstly, in the betting industry, where more accurate predictions can influence betting odds, fostering a more engaging and dynamic betting environment. Secondly, in player analysis and scouting, enhanced prediction accuracy aids teams in identifying players with optimal performance potential, contributing to more strategic player recruitment and development.

The primary focus, however, is on how these improved models can directly benefit player performance. By providing more precise data, coaches can develop targeted training programs and tactical decisions, leading to a higher overall shot accuracy. This advancement in player performance has a domino effect, enhancing team success and fan engagement. These improvements, even if incremental, can lead to significant competitive advantages, manifesting in increased revenue streams such as ticket sales, merchandise, and sponsorships.

To contextualize these implications, we might consider a scenario with a team like the Phoenix Suns. An illustrative 1% increase in their shooting accuracy could potentially yield more points per game, possibly turning tight matches into wins. This heightened performance not only bolsters fan experiences but also impacts the team's financial health, from game-day earnings to broadcasting rights. In the fiercely competitive realm of professional sports, these gains, though seemingly small, are essential for long-term financial sustainability and success.

1.4 Thesis Structure

This thesis begins with a background overview, research objectives, and potential contributions. The literature review section then explores predictive models in analytics, focusing on their evolution and importance. It specifically addresses basketball analytics, examining changes in shot strategies and the role of ML in shot prediction, concluding with a discussion on research gaps and future directions. Chapter 3 provides a detailed statistical analysis of the data used. In the Methodology section, the thesis outlines the approaches for Ensemble Methods, detailing data preprocessing, feature selection, and model architecture for Random Forest and XGBoost, followed by a segment on Feedforward and Recurrent Neural Networks (FNNs and RNNs). The Discussion section interprets findings, considering their theoretical and practical implications. The

thesis wraps up with a summary of key findings and potential future directions in basketball analytics research. A structured visual representation of the thesis is available in Figure 1.

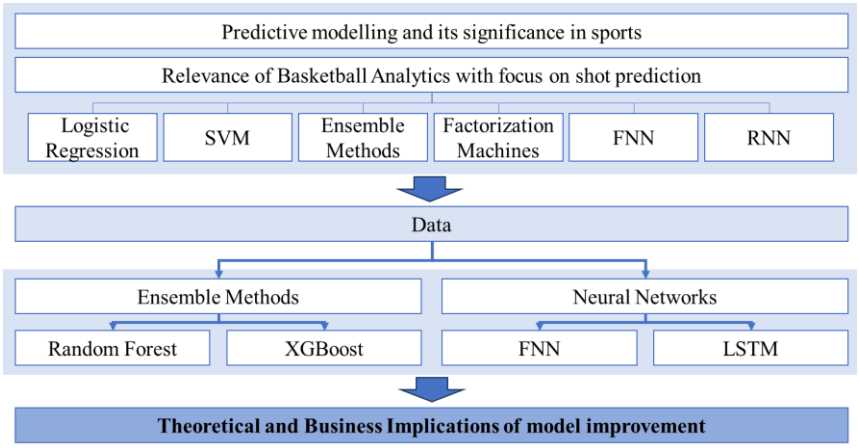

Figure 1: Visualization of Thesis structure

2. Literature Review

This literature review delves into the transformative impact of predictive models and ML in sports analytics, reshaping how games are strategized and analyzed. It begins by exploring predictive models in a broad context, focusing on their evolution and newfound capabilities in sports. The review then shifts to basketball analytics, emphasizing the changes in NBA shot strategies and the growing dependence on ML for shot prediction. The section concludes with a thorough analysis of existing studies on NBA shot success prediction, pinpointing crucial research gaps and unexplored areas. These identified gaps set the stage for this thesis, steering our research towards significant contributions in basketball analytics.

2.1 Overview of Predictive Models

Predictive modelling is a branch of analytics and statistical methods (Kumar 2018). It is "the process by which a model is created or chosen to try to best predict the probability of an outcome" as defined by Geisser (1993). By evaluating current and historical data and applying methods from artificial intelligence, ML, statistics, and data mining, predictive modelling is used to forecast

future events (Elkan 2013). Predictive models describe the relationship between different attributes and their performance, when assessing the probability that a comparable unit in a distinct sample shows a particular outcome (Kumar 2018). There is a wide range of predictive models with varying strengths and limitations in predictive modelling. Generally, they are grouped into two categories: regression models and classification models. Regression models forecast a numeric response, whereas classification models predict whether given values belong to a particular class (Kuhn and Johnson 2019).

Predictive modelling originated in classical statistics, where statistical models were used to forecast events in the future. With developing techniques in the field, it has become essential for organizations and has grown to be one of the most critical researched topics (Siegel 2016). Different reasons led to this: data is becoming more available in size and categories, computers and software are more accessible for users, and developments in ML algorithms and artificial neural networks are reforming predictive modelling. To stay competitive, organizations need to adapt to the use of predictive modelling (Kumar 2018), which can be observed in a variety of different domains: In **retail**, it is used for demand prediction and assessing media campaign profitability due to enhanced data sources and quality (Bradlow et al. 2017; Pereira 2021; Giri et al. 2019) and in the **political** sphere, especially in international politics, predictive modeling offers insights where controlled experiments are impractical (Cranmer and Desmarais 2017). In **medicine and healthcare**, it aids in predicting patient outcomes, assisting clinicians in decision-making across diverse areas like brain death and diabetes implications (Alanzani et al. 2017; Labarère et al. 2014; Liu et al. 2011; Fregoso-Aparicio 2021). In **financial services**, investors leverage predictive models for making informed decisions on asset prices and financial distress (Alhnaity and Abbod 2020; Chen and Du 2009).

In summary, predictive modeling has emerged as a cornerstone in a multitude of industries, underpinned by the exponential growth of data and advancements in computational capabilities. From retail and politics to healthcare and financial services, its application demonstrates a significant impact on decision-making processes, strategy formulation, and future projections.

2.2 The Emergence and Rise of Machine Learning in Sports Analytics

As the sports market continues to expand, driven by technological advancements, the field of sports analytics has significantly emerged. This growth reflects the increasing competitiveness in professional sports (Global Sports Market Revenue 2023). Sport Analytics constitutes the field of statistics involving the collection, processing, and interpretation of data to make decisions for a competitive advantage in sports. Sports Analytics has a wide range of applications such as player's success analysis, game strategies, rules feasibility, score prediction or player's health conditions (Minusha 2016). These insights can be gained through different data modelling techniques and optimizations with the objective of reaching valuable recommendations (Sarlis et al. 2020).

Traditionally, success in sport was mostly based on superior ownership of players, coaching and front offices with decisions based on past games or gut feeling. However, starting in the early 2000's within Baseball, statistics and data began to be used to make strategic decisions modernizing and revolutionizing the world of sports. As of today, every professional team largely relies on analytics for its competitive advantage (Steinberg 2015). And with the substantial growth in the volume, quality, and accuracy of data in recent years, there has been a notable enhancement in sports performance resulting from increasingly data-driven strategies (Krebs 2022). Not only has Sports Analytics allowed teams and players to reach more efficient and strategic decisions but it has also attracted fans and supporters. To gain a better understanding of games or seasons, fans

have also begun to consume analytical content through sport analytics conferences, research websites or platforms with predictive models (Steinberg 2015).

The emergence of predictive modeling in sports analytics is closely tied to the advancements in ML through the training of algorithms on vast amounts of data to make predictions. These algorithms learn patterns from historical data, enabling them to make informed predictions on new, unseen data (Thabtah 2019).

Over the past few decades, key studies have consistently demonstrated the effectiveness of ML in various applications in sports. These include game activities, with models focusing on match outcomes, ball tracking, shot classification, and sports betting. Another common application of ML in this industry is in talent identification and acquisition, which involves recruiting players and analyzing their performance. Moreover, training and coaching have also benefited from ML, particularly in studying the efficacy of tactical planning, injury modeling, and team formation assessment. Additionally, ML is employed in fan and business-focused applications, such as measuring a player's economic value, predicting event attendance, and optimizing ticket pricing (Beal et al. 2019).

ML algorithms are broadly categorized into two main types: supervised and unsupervised learning. These classifications are based on how the algorithms process and learn from data to make predictions or identify patterns. In the context of sport analytics, the problem dictates what type of ML algorithm is deployed (Chmait et al. 2021). A specialized subset of ML, DL, has garnered significant attention in recent years. It employs neural networks with many layers to analyze various factors of data. Utilizing neural networks that mimic the human brain's structure, vast amounts of data can be processed and capture intricate patterns across multiple layers of computation. For basketball shot prediction, this means analyzing game footage to understand

player movements, ball trajectories, and even facial expressions, offering insights that were previously unimaginable (NCBI 2022).

The paper "The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review" by Rory Bunker and Teo Susnjak conducts an extensive review of ML model applications in team sports, covering research from 1996 to 2019. It focuses on evaluating the most popular algorithms and assessing the accuracy of sports outcomes (Bunker et al. 2019). Bunker and colleagues (2019) highlight that diverse training-validation data splits, extensive feature engineering, and the use of distinct feature subsets have contributed to more successful predictions in sports. However, they also acknowledge challenges in prediction accuracy across different sports, attributing this to varying scoring systems and competition structures. For instance, sports with low scoring, such as football, may yield less accurate results (Bunker et al. 2019).

In conclusion, AI is poised to continue transforming how we play, follow, and analyze sports (Chmait et al. 2021). The increasing volume of recorded data, coupled with its growing availability, is set to drive the adoption of big data technologies more broadly within the industry. This is essential for teams aiming to maintain a competitive edge (Banerjee 2023). Such advancements in sports analytics are expected to lead to enhanced game performances (Rein et al. 2016), improved fan experiences, and reduced risk of injuries (Banerjee 2023).

2.3 Predictive Modeling in Basketball

The realm of basketball has undergone a significant transformation, paralleling trends observed in other sports. This evolution, as outlined by Steinberg (2015), has revolutionized both game strategy and player performance. By harnessing the power of both current and historical data, basketball teams and analysts have developed various methods for forecasting and optimizing team performance, thereby enhancing decision-making processes (Sarlis et al. 2020).

Specifically, the advancement and increased collection of data have significantly enriched the scope of predictive modeling in basketball. Loeffelholz et al. (2009) utilized ML techniques to predict NBA match outcomes, focusing on vital statistics such as field goal percentage, three-point percentage, and free-throw percentage. Their study underscores the importance of these metrics, revealing that averages of these features within the current season are highly indicative of future game performances, thus enhancing the accuracy of predictions for upcoming matches (Loeffelholz et al. 2009). Similarly, Zdravevski and Kulakov (2009) explored the efficacy of algorithms in predicting NBA match outcomes, achieving a notable classification accuracy of 72.8%. These studies highlight the shift in focus within basketball analytics - from traditional statistics to more nuanced and predictive data points that include shooting efficiencies, player movements, and in-game situational variables.

Along with the emergence of predictive modeling techniques in basketball, advanced data collection technologies are increasingly implemented in the sport. NBA teams are currently using a "player tracking" technology where team efficiency is measured through player movement (Steinberg 2015). This tracking is possible thanks to their software "SportVU" with 6 installed cameras in the catwalks of arenas to record all the movements of basketball players in games. The optical tracking system has cameras that capture spatial positions 25 times every second measuring speed, distance, player separation and ball possession (Shah and Romijnders, 2016). Thanks to these data records, in-depth investigations into various aspects of the game, including defensive contributions can be made (Minusha 2016). Crucially, there are some concerns within the dataset such as irregular sparse data because many players change leagues or teams very recently as well as short careers (Sarlis et al. 2020). Another concern is the difficulty to distinguish dominant and opponents' performance simultaneously (Sarlis et al. 2020).

2.4 Shot Prediction in Basketball

While various performance indicators contribute to the success of a basketball team - ranging from coaching strategies to player fitness (Zhang et al. 2019; Huyghe et al. 2022) - one factor stands out as particularly crucial: shooting accuracy. As noted by Akers et al. (1992), García et al. (2013), and Çene (2018), the accuracy of field goals is a decisive element in determining the outcome of basketball games across professional leagues. This observation is further supported by Wang and Zheng (2022), who reported temporal variations in field goal accuracy based on shooting distances. As Teramoto and Cross (2010) and Mikolajec et al. (2013) have highlighted, both offensive and defensive efficiencies are vital for a team's success. However, as supported by the studies cited above, it is the precision in shooting that often tips the scales. Therefore, the evolution and prediction of shooting patterns have become increasingly significant within the realm of basketball analytics, warranting a closer examination.

The landscape of shooting in the NBA has undergone a significant transformation over the past decade, most notably characterized by the decline of the mid-range jump shot. Teams and players are increasingly focusing on either taking three-pointers or driving to the basket for layups, thereby marginalizing the mid-range game (Wang and Zheng 2022). According to Wang and Zheng (2022), the percentage of field goals taken from beyond 24 feet has nearly doubled since the 2011–2012 season. This observation is further supported by Vicent et al. (2022), who note that the average number of 3-point shots per game has more than doubled over the last 22 seasons. This shift began to crystallize around the 2012-2013 NBA season, marking the advent of what is now widely known as the "three-point revolution" (Kilcoyne 2020).

Crucially, one of the most influential factors behind this shift is the rise of data analytics in basketball, which has brought to light the higher expected value of three-point shots compared to

mid-range shots by quantifying their efficiency and impact on game outcomes. Kilcoyne (2020) highlights the role of data analytics in shaping modern NBA strategies, stating that nearly every NBA team has hired data analysts to work with coaches and front office staff.

Daryl Morey, former General Manager of the Houston Rockets, is often credited as a pioneer in applying analytics to basketball strategy. His data-driven approach, commonly referred to as "Moreyball," led the Rockets to focus on taking highly efficient shots, predominantly three-pointers and layups, effectively abandoning the mid-range game (Kilcoyne 2020; Vicent et al. 2022). The analytical shift has also influenced the pace of the game. Kilcoyne (2020) notes an increase in the number of possessions a team has per game, indicating a faster pace. This increase in speed has contributed to more transition opportunities, which are often capitalized on through analytics-driven shot selection, such as opting for a layup or finding an open shooter. This strategic focus has, in turn, improved shooting efficiency for both two-point and three-point attempts (Vicent et al. 2021).

2.5 The Strategic Importance of Shot Prediction

Building on the transformative role of data analytics in basketball, underscored by Kilcoyne (2020), and the growing focus on shot efficiency, the strategic significance of shot prediction has become increasingly clear. This advanced approach to analytics ensures that strategic decisions are well-informed and tailored to the flow of the game, enhancing performance when it matters most (Li and Feng 2020). This is supported by research of Wang and Zemel (2021) which adds a critical dimension to this understanding. They found that shooting accuracy for distinct positions at different shooting distances has generally increased in the last five years. Notably, there have been increases in the accuracy of field goals with shooting distances over 24 feet for players in the roles of center, power forward, and shooting guard. This trend indicates a shift in player capabilities and

strategic play, further underscoring the role of data analytics in enhancing basketball strategies and player performance (Wang and Zemel 2021).

The predictive insights of AI models thus enable players and coaches to make informed decisions about when to take a shot or to create an opportunity for a higher-quality attempt (Li and Feng 2020). In other words, the moment when a player decides whether to shoot or hold for a better chance is dictated by predictive analytics. This complex decision-making process has been enhanced by DL models that offer a more nuanced understanding of player positioning and game dynamics, going beyond traditional hand-crafted features (Harmon et al. 2021).

Advancements in technology allow for real-time feedback on shot success, leading to rapid improvements in shooting accuracy and performance. This is enhanced by vision sensors and trajectory learning algorithms, which give players instantaneous feedback, facilitating the correction of shooting techniques and reinforcing successful patterns (Smith and Singh 2023). ML models not only predict shot success but also identify weaknesses in a player's shooting form, allowing for targeted improvement. By analyzing shooting patterns and outcomes, coaches can pinpoint specific areas where a player may struggle, such as long-range shots or angles on the court (Nakai et al. 2023). Analytics not only informs current strategies but also anticipates future shifts, ensuring that the endeavor of predictive modeling remains an essential, ever-evolving component in the search for competitive advantage in professional basketball.

2.6 Models Predicting Shot Success in NBA

This section presents a comprehensive literature review focused on the utilization of ML and DL models in predicting shot success in basketball. It delves into various studies that have applied these advanced technological approaches specifically within the NBA context. An overview of all the studies considered can be found in Appendix A.

Logistic Regression

Logistic regression, commonly used in binary classification problems like predicting basketball shot success, models the probability of an outcome using a logistic function that assigns a value between 0 and 1, indicating the likelihood of a shot being successful or not. It considers various predictor variables, such as the shooter's position and temporal variables, which are weighted based on their influence on the outcome (Meehan 2017).

Meehan's (2017) approach to logistic regression was based on a NBA 2014-2015 dataset comprising around 42,000 samples, which led to a prediction accuracy of 59%. While this figure is notably higher than his baseline accuracy of 50% — which represents random guessing — it wasn't particularly remarkable. This outcome could be attributed to several factors, including the removal of many categorical features, as suggested by Meehan himself. The greatest limitation of logistic regression constitutes its assumption of a linear relationship between variables and the outcome, potentially oversimplifying the complex dynamics of a basketball game.

Support Vector Machines (SVM)

Support Vector Machines (SVMs) have been a cornerstone in the ML community, particularly for classification tasks. Their ability to handle high-dimensional data by leveraging the kernel trick makes them a promising candidate for complex prediction tasks, including predicting the success of basketball shots (Meehan 2017). Meehan hypothesized that SVMs, with their feature-generating kernel trick, would outperform other models in terms of accuracy and bias reduction (Meehan 2017). However, his findings revealed that the SVM model reached a peak accuracy of only 55% on a dataset of 40,000 training examples. This was notably lower than anticipated, suggesting that the additional features generated by the kernel trick did not significantly reduce bias in this context.

One possible explanation for this outcome, as Meehan points out, could be the general preprocessing steps taken before model training. He removed categorical variables with more than two categories, such as game id, matchup, and closest defender, among others. While this preprocessing was deemed necessary for logistic regression, its application to SVMs might have inadvertently compromised the model's performance (Meehan 2017).

Random Forest

In the realm of basketball data analysis, numerous studies have emphasized the application of Random Forest techniques. Random Forest operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. This algorithm is known for its robustness and ability to handle large datasets with high-dimensional features (Meehan 2017).

Oughali et al. (2020) created a Random Forest model to predict shot success leveraging a dataset encompassing 203,591 shots from the NBA 2014-2015 regular season. The Random Forest model was trained using four specific attributes as predictors: the shot clock, the number of dribbles, the shot distance, and the closest defender’s distance attributes. The research achieved an accuracy of 57% using the Random Forest model. Notably, Brett Meehan's research also utilized data from the same NBA 2014-2015 season, drawing parallels with Oughali's study. Using 42.105 shot examples, Meehan's study achieved a peak accuracy of 61% with the Random Forest algorithm. Additionally, the study identified the relative importance of various features. In order of increasing importance, these features were dribbles, period, final margin, touch time, shot clock, closest defender distance and shot distance (Meehan 2017).

Interestingly, while Oughali used a larger dataset (203,591 shots), Meehan's study, with only a fifth of Oughali’s datapoints (42.105), achieved higher accuracy (61% vs. 57%). This could be due to

Matlab's TreeBagger, used by Meehan for extensive hyperparameter tuning, allowing for a more optimized model fit. In addition, Meehan utilized a broader set of features, including period, final margin, touch time, shot clock, closest defender distance, and shot distance, compared to Oughali's selection of four features: shot clock, dribbles, shot distance, and closest defender's distance (Meehan 2017; Oughali et al. 2020). Consequently, it can be argued that the addition of features 'period' and 'final margin' offer a more detailed representation of the situation surrounding each shot, allowing the model to capture more nuances and thereby improve its predictions. Crucially, based on the feature importance analyses conducted by Meehan and Oughali et al. it can be concluded that the shot clock, the number of dribbles, the shot distance, and the closest defender's distance constitute important attributes in determining shot success (Meehan 2017; Oughali et al. 2020).

XGBoost

As discussed previously, XGBoost also constitutes an ensemble method, however, compared to the Random Forest model it has distinct operational mechanisms. While Random Forest capitalizes on the diversity of its trees to achieve robustness and accuracy, Gradient Boosting and XGBoost focus on reducing both bias and variance by correcting errors iteratively. In other words, while Random Forest trains each tree independently using random data samples, XGBoost builds trees sequentially, correcting errors from previous trees.

Following the literature, Shah and Romijnders delve into the application of Gradient Boosting, achieving an AUC of 0.719 which can be interpreted in the context of how well the model distinguishes between two classes (Shah and Romijnders 2016). Separately, Meehan's paper corroborates the effectiveness of boosting in predicting shot success. Meehan's XGBoost achieves a 68% accuracy rate after parameter tuning, hereby constituting the best model of his comparative

analysis (Meehan 2017). Meehan suggests that boosting, especially XGBoost, is a robust method due to its ability to aggregate multiple weak learners into a strong predictor, thus reducing both variance and bias in predictions. In Meehan's XGBoost model, spatial/distance data was found to be the most predictive of shot success, and therefore served as primary predictors (Meehan 2017).

Finally, Oughali et al. (2020) achieved an equal accuracy score on their XGboost of 68%. Similar to Meehan's study, the XGBoost model is employed to analyze a dataset from the 2014-2015 NBA season containing over 200,000 shots. Initial models included four features, but further analysis expanded this to seven, with significant emphasis on the four most predictive ones. The model's effectiveness is enhanced by employing the GridSearchCV function for parameter tuning (Oughali et al., 2019). Interestingly, like the study of Shah and Romijnders (2016), they already achieve a significantly high score of 60% on their non-fine-tuned XGboost (Oughali et al., 2019). Hence, we can conclude that even without extensive parameter optimization, XGBoost models demonstrate a strong baseline performance in predictive tasks.

Factorization Machines (FM)

In the study titled "Shot Recommender System for NBA Coaches," Wright and colleagues delved into the potential of the FM model for predicting NBA shot outcomes using data from the 2015-2016 season (Wright et al. 2016). They argue that traditional methods like logistic regression or support vector machines often falter with the NBA's sparse datasets, given that many player-style-location combinations might not be present in the training data. Wright et al. championed the FM model for its adept handling of sparse data and its scalability to vast datasets. They highlighted the model's ability to provide latent factors that encapsulate player preferences and shot features. To assess Wright et al. (2016) opt to provide the RMSE rather than the accuracy metric. Relying solely on RMSE has its limitations, particularly when comparing with studies that use accuracy metrics.

This is because RMSE, a measure typically used for regression problems, quantifies the average magnitude of errors, making it less intuitive for classification tasks like shot success prediction, where accuracy, representing the proportion of correctly classified instances, provides a more direct assessment of performance.

Feedforward Neural Network (FNN)

The Feedforward Neural Network (FNN) is a fundamental and straightforward structure in the neural network family. Neural networks consist of layers of interconnected nodes or 'neurons', each capable of performing simple calculations. In FNNs, these layers are arranged linearly, where each layer's output becomes the input for the next layer, forming a 'feedforward' architecture. The Feedforward Neural Network (FNN) becomes a Multi-Layer Perceptron (MLP) when it includes one or more hidden layers in addition to the input and output layers, allowing it to handle more complex data relationships beyond what a single-layer perceptron can manage (Meehan 2017).

The comparison of the studies by Meehan and Harmon et al. offers a window into the varied applications of Feedforward Neural Networks (FNNs) in predicting NBA shot success. Meehan's (2017) FNN relied on the same dataset leveraged in his baseline studies, namely the NBA 2014-2015 season consisting of 122,502 examples of shots.

Meehan tested two one-layer FNNs, both with 50 units in their hidden layers, but with different activation functions. The first network used the sigmoid activation function for both the hidden and output layers. In contrast, the second network utilized the RELU activation function for the hidden layer and the Sigmoid function for the output layer, with both networks incorporating L2 regularization to avoid overfitting. Using Log Loss as the cost function (Figure 2), the entirely Sigmoid neural network achieved an accuracy of 55% on the test set. The RELU/Sigmoid network, however, recorded a slightly lower accuracy of 53% on the training development set but matched

the 55% accuracy on the test set. As hypothesized by Meehan himself, a significant limitation of the study includes the removal of categorical data and a scarcity of spatial data. Adding more diverse and spatially detailed data to the input vectors could lead to more effective shot predictors.

$$\frac{1}{m}\sum_{i=1}^{m} -y^{(i)}\log p^{(i)} - (1-y^{(i)})\log(1-p^{(i)})$$

Figure 2: Log Cost Function (Meehan 2017)

Harmon et al. (2016) uniquely relied on the dynamic SportVU data from the 2012-2013 NBA season. This allowed them to include critical on-court dynamics such as player and ball positions at the moment of the shot, player speeds over the last five seconds, ball speed, and other spatial relationships on the court. Another dimension added to the analysis is the ball possession time for each offensive player and the count of all individuals, including teammates, near the shooter (Harmon et al., 2016). In terms of the model's architecture, Harmon et al. constructed a three-layer FNN, incorporating a Softmax function at the output layer. The Error Rate and Log Loss score on the test and validation set for Harmon et al.'s FNN can be viewed in Figure 3. As can be seen, the error rate is approximately 55%, implying that the model incorrectly predicts the outcome of a shot over half the time (Harmon et al. 2016). This high error rate may suggest limitations in the model's ability to capture the complexities of shot success in basketball or could be indicative of a need for more nuanced features or a larger dataset for training. The log loss on the test set is equal to 0.807, which indicates that while the model does provide better-than-random predictions, there is significant room for improvement in terms of the confidence and accuracy of its output.

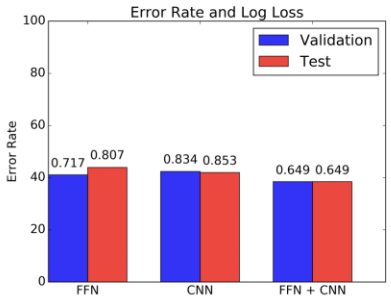

Figure 3: Error Rate and Log Loss (Harmon et al., 2016)

A relatively low accuracy of Harmon et al.'s Feedforward Neural Network (FNN) compared to Meehan's study, despite their access to the detailed SportVU dataset, may be attributed to several factors. Firstly, the complexity of their three-layer FNN could lead to overfitting, where the model excessively adapts to the training data, hindering its performance on new data. Secondly, the rich yet intricate nature of the SportVU data, encompassing dynamic and spatial features, might introduce higher complexity and noise, potentially overwhelming the model. Lastly, the specific architecture and training approach of the FNN, including choices of activation functions, normalization methods, and optimization algorithms, might not have been optimally tuned for the dataset, impacting learning and prediction accuracy.

In conclusion, the exploration of Feedforward Neural Networks (FNNs) in the studies by Meehan and Harmon et al. reveals some critical insights and limitations in the application of advanced DL techniques. Interestingly, Meehan's FNNs did not achieve significantly higher accuracy or performance compared to his simpler baseline models. And despite the anticipation that Harmon et al.'s use of dynamic SportVU data would lead to a highly accurate FNN, their model's accuracy was lower than expected. This outcome highlights the challenges in neural network applications, where increased data complexity and model architecture do not always translate into better predictive performance, emphasizing the need for a balanced approach in model design and feature selection in sports analytics.

Recurrent Neural Networks (RNN)

Transitioning from the simplicity of FNNs, Recurrent Neural Networks (RNNs) introduce a layer of complexity with their ability to handle sequences. Their success in domains like text, music, and motion data underscores their potential in capturing the temporal dynamics of a player's movement and shot technique, offering a more nuanced prediction of shot success (Yan et al. nd). Similarly, RNNs also seem to be well-suited for processing spatial data. A study by Wang and Zemel showcased the potential of RNNs in classifying NBA offensive plays, indicating that RNNs can extract intricate relationships from spatial data (Wang and Zemel 2016). Although their primary focus was on play call classification rather than direct shot prediction, the methodology is closely related.

By processing SportVU tracking data, they demonstrated the potential of RNNs to extract relationships from spatial data. A crucial component of the research is the incorporation of LSTM units into the RNNs, since LSTMs aid in getting around the drawbacks of conventional RNNs, especially when it comes to learning long-term dependencies (Wang and Zemel 2016). This is essential to comprehend basketball's play sequences. Their results, underscore the potential of adding more spatial data to improve prediction accuracy. The study's results demonstrating that RNNs, and particularly those with LSTM units, significantly outperformed baseline models in classifying offensive plays. The RNN model achieved a top-1 classification accuracy of 66% (Wang and Zemel 2016).

Another study by Shah and Romijnders, is more pertinent to the scope of this thesis, given its direct focus on the objective of predicting shot success (Shah and Romijnders 2016). Their study employs a two-layered LSTM RNN, which notably achieved an AUC of 0.843. This performance significantly surpasses that of traditional models, which recorded AUC scores of 0.558 and 0.719

obtained by a general linear model and gradient boosted machines, respectively (Shah and Romijnders 2016). The RNN's performance is highlighted when predicting a make or miss using half a second of data with the ball 8 feet away from the basket. Hence, these results suggest that the sequential RNN models can effectively learn and generalize nonlinear behavior from sequential data. However, an obvious limitation is that the narrowed focus (a narrow scope of shots from 8 feet away) means the findings are less applicable to other types of basketball shots (Shah and Romijnders 2016).

In conclusion, the exploration of Recurrent Neural Networks (RNNs), especially those employing Long Short-Term Memory (LSTM) units, in the realm of basketball analytics, highlights their significant potential in enhancing shot prediction accuracy. While the primary objective of the Wang and Zemel study was not directly aligned with shot prediction their study demonstrates the effectiveness of RNNs in capturing complex sequential and spatial patterns in basketball data, an approach that can be adeptly applied to shot prediction. Further, the study by Shah and Romijnders distinctly establishes the prowess of RNNs in the context of shot success prediction, albeit in the scope of 8 feet shots. Their LSTM RNN model, achieving high accuracy without extensive hyperparameter tuning, showcases the innate capability of these networks to learn and generalize from sequential data, setting a promising precedent for future research and applications in this field.

2.7 Research Gaps and Limitations

The current landscape of utilizing ML and DL models for predicting shot success in basketball, as reviewed in the preceding sections, reveals significant advancements as well as notable gaps and limitations in existing research. Across the literature, we have seen that Ensemble methods including Random Forest and XGBoost, are consistently well-performing and robust, yielding the highest accuracy scores. Crucially, we have seen that Factorization Machines (FMs) known for

their proficiency in sparse datasets, face evaluation challenges due to Wright et al.'s reliance on RMSE, which hampers direct comparability with other models assessed using accuracy metrics. This inconsistency in evaluation criteria leads to the exclusion of FMs from our focus, as we aim for a more standardized approach to model assessment.

Despite their robustness, ensemble methods typically achieve accuracy levels that do not surpass the 70% mark. This limitation reflects the inherent unpredictability and complexity inherent in basketball, a sport where a myriad of variables influences each shot. Achieving exceptionally high accuracies, such as 90-95%, seems unrealistic considering the nuanced nature of basketball shots made by humans. Factors as subtle as a player's elbow or foot positioning, or even a slight imbalance, can significantly alter the shot's outcome (Meehan 2017). These details, often not captured by available features, play a crucial role in determining whether a shot is successful, further compounding the challenge of achieving extremely high prediction accuracies in this context. Nevertheless, we suggest there is still untapped potential for the utilization of ensemble methods in NBA shot prediction which largely resides in increasing the scope of the data.

We propose that by expanding the dataset to include multiple seasons, the models can access a richer, more varied pool of data, enhancing their ability to learn and predict with greater accuracy. This is confirmed by Shah and Romijnder's (2020) study who established that increased training data correlated with higher AUC scores in their findings. They suggest expanding the dataset not only enriches the learning depth but also enables the network to recognize and adapt to a broader array of patterns in shot trajectories (Shah and Romijnders 2016). For instance, a larger data scope may allow to isolate the effects of individual players or team strategies.

Similarly, we suggest that the limited scope of data has significantly constrained the performance of neural networks we discussed, specifically, the FNN and RNN. Neural networks, renowned for

their ability to process and learn from vast amounts of data, are especially sensitive to the size of the dataset. The utilization of data from only one season significantly restricts the neural network's ability to fully exploit its potential for learning and pattern recognition (Shah and Romijnders 2016).

Apart from a limited data scope, we have identified specific limitations in existing studies that leverage FNNs and RNNs in predicting shot success that this study can built on. Regarding the FNN neural network, we recognize its potential based on the study conducted by Meehan. Meehan's (2017) implementation of an FNN network with a single 50-node layer achieves a modest accuracy of 55%. The limitations of this study, including the removal of categorical data and a scarcity of spatial data, suggest that enriching the input vectors with more diverse and spatially detailed data could lead to more effective shot predictors (Meehan 2017).

With respect to RNNs, the potential is underscored by the work of Shah and Romijnders (2016) who utilized a two-layered LSTM model, achieving a considerable AUC-ROC score of 0.843. Despite only using positional data and default hyperparameters, their RNN outperformed feature engineered GBM models (Shah and Romijnders 2016). Strategic optimization of hyperparameters and more nuanced feature engineering are expected to unlock the full potential of these networks, which have been somewhat underutilized due to default parameter settings. Nevertheless, Shah and Romijnders convincingly validated the utility of RNNs in learning from and interpreting sequential data in a basketball context.

2.8 Discussion

Based on the potential and limitations outlined in the literature this study aims to build upon the deployment of the Random Forest model, XGBoost, FNN and LSTM in predicting NBA shot success. We expect to enhance predictive modeling of shot success in three ways.

1) Increasing the data points by including 5 seasons (NBA 2019-2023) should enhance the learning capacity of the model, leading to greater accuracy (especially for neural networks). In addition, using data over multiple seasons allows the model to detect other patterns such as the effect of individual players/teams on shot accuracy.

2) Incorporating more categorical data can provide a more holistic view of player actions and shot contexts, potentially refining the model's predictions.

3) Optimization of hyperparameters and strategic feature engineering could harness the full capabilities of machine learning models, which has been largely underutilized due to a reliance on default parameter settings in existing studies.

As we move forward with our research, the establishment of two baseline models becomes a crucial initial step. These models, namely a Random Forest model and an XGBoost model, have been chosen based on their recognition in the literature as state-of-the-art methodologies known for their high performance and reliability. The deployment of these models serves a dual purpose. First, they provide a robust foundation against which the performance of our proposed neural network models can be rigorously compared and evaluated. Secondly, these models play a key role in identifying the most predictive features, which can subsequently be utilized as essential inputs in the construction of neural networks. Following the establishment of these ensemble models, our next step will be to delve into the further potential of the identified neural networks, specifically FNN and RNN. This exploration aims to make significant contributions to the research gaps we have identified.

To conclude, by harnessing the potential of machine learning models and addressing the challenges and limitations identified in previous studies - the scope of data points, categorical data, and optimization of model parameters - we anticipate providing valuable advancements in the

analytical techniques used in the sport, thus offering both theoretical and practical implications for teams and analysts alike.

3. Data and Context

3.1 Introduction to the Data

The dataset used for our research is an NBA (National Basketball Association) regular season shot location dataset spanning from the 2003-2004 season to the 2022-2023 season. However due to computational and processing limitations the study will focus on the 5 most recent seasons (2019-2023). The dataset was proposed as one of the field lab topics for Business Analytics, making it an academically relevant and well-structured dataset for research purposes. The dataset was made available through a GitHub repository maintained by Dominic Samangy, a Basketball Analytics Coordinator for the New Orleans Pelicans. This repository provides useful additional resources to help with data exploration and visualization. However, the original source of the data is from the NBA.com website, the official platform for NBA statistics.

The dataset's extensive timeframe of games allows for the possibility of studying long-term trends and changes in shot success rates, player performance, and game dynamics over the years. Furthermore, by incorporating data from several seasons, it's possible to analyze how the NBA's evolving strategies, rule changes, and player abilities have influenced shot outcomes. However, as previously mentioned, the study in question will focus on the 5 most recent seasons due to computational power limitations and model optimization.

Geographically, the dataset focuses on the NBA premier basketball league, covering games played across various cities and arenas in Canada and United States. Given that the NBA is followed globally, the geographical scope of the dataset aligns with the league's reach and impact. Moreover,

the dataset offers insights into the spatial and temporal dimensions of the game at a granular level, enabling precise analysis of shot success factors.

3.2 Data Descriptions and Statistics

Given that our research focuses on a timeframe of five seasons, the total size of our dataset is 212.8 MB. The datafiles were provided in a csv format and include a total of 26 features which could be divided into three sections: player and team information, game information and shot characteristics. Regarding player and team information, the dataset includes useful variables such as player's or team's name and id's as well as player's positions which are useful to help analyze how player positions, team strategies, and individual player skills influence shot success. The dataset also includes features with game information such as the name of the home team, away team and the season or date of the event. These variables give us game context which could help us analyze shot success rates could be impacted by contextual game situations such as home-court advantage, the specific season or the date of the game.

Variables on shot characteristics are significantly impactful and therefore constitute our focus for building a predictive model on shot success. These features collect characteristics of each shot attempt such as location coordinates, hoop distance, time remaining, game situation, type of shot or court zone. Indeed, the characteristics of each shot attempt are crucial in predicting the success rate of a shot. Within this section of shot characteristics lies the target variable for our shot prediction model, a binary variable determining whether the shot was made or missed. The different variables mentioned within the NBA dataset each display a variety of data types, as detailed in Appendix B. Regarding the number of instances, the dataset counts with 1.032.499 rows, each representing a shot attempt from a player.

The dataset, enriched with 1,027,134 entries post-cleaning, encompasses a diverse range of variables that paint a detailed picture of basketball shooting trends. A critical component of the dataset is the SHOT_DISTANCE, which varies from a mere 0 feet to a remarkable 88 feet, averaging around 13.5 feet. Intricately linked to the spatial analysis of shots are the LOC_X and LOC_Y variables, representing the coordinates of each shot attempt. Another temporal dimension is added by the QUARTER variable, which unusually ranges from 1 to 8, likely accounting for overtime periods in games. Additionally, the MINS_LEFT and SECS_LEFT variables provide insights into the critical moments of the game when shots are taken, highlighting the pressure-packed nature of basketball.

The dataset also includes several categorical variables, offering a deeper understanding of the context in which shots are taken. The TEAM_NAME and PLAYER_NAME variables underscore the variety in the dataset, with 31 teams and 948 players represented, showing the wide participation across the basketball landscape. The dominance of guards, especially shooting guards, in the POSITION variables reflects modern basketball's emphasis on perimeter play. The SHOT_MADE variable correlates closely, with misses being more frequent, a testament to the challenging nature of scoring in basketball. The ACTION_TYPE variable, with its 48 unique shot types, with 'Jump Shot' being predominant, adds another layer of complexity to the analysis. The SHOT_TYPE is crucial in distinguishing between two-pointers and three-pointers, crucial for understanding modern basketball's scoring dynamics. Lastly, ZONE_NAME and ZONE_RANGE, detail the court areas and distance categories from which shots are taken, revealing preferred spots and strategies employed by players.

Shot Success Distribution

The statistical analysis of this comprehensive dataset offers insights into the dynamics of basketball shooting. A balanced representation of shot outcomes, with approximately 53.53% misses and 46.47% makes, provides a solid foundation for unbiased predictive modeling, as can be seen in Figure 4. This balance is crucial in avoiding any skew towards predicting one outcome over the other.

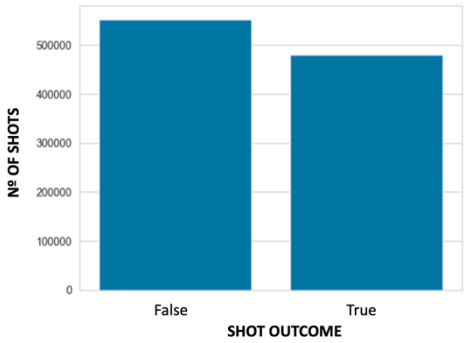

Figure 4: Distribution of Shot Outcomes

Shot Location Distribution

Comparing two-point and three-point field goals, the data aligns with conventional basketball wisdom. Two-pointers, generally taken closer to the basket, exhibit a higher success rate compared to the more challenging three-pointers. Figure 5 depicts a distribution of all shot locations. Regarding the descriptive statistics of shot coordinates, LOC_X has a mean very close to 0, which makes sense since the court is symmetrical, and shots are taken from both sides. LOC_Y has a positive mean, indicating that shots are taken from a range of distances from the basket, with the majority being closer than the three-point line. The standard deviation for both LOC_X and LOC_Y shows that there is a wide spread of shot locations across the court. The minimum and maximum values for LOC_X are within a typical range for an NBA court, while the maximum for LOC_Y

indicates some very long shots. Figure 5 shows a visual representation of the coordinates from each shot made where the comparison of two-point and three-point field goals is clearly appreciable.

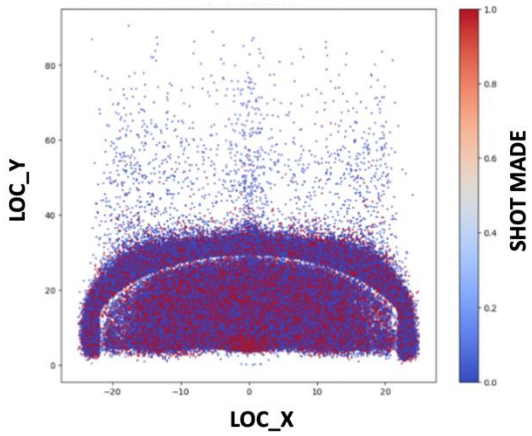

Figure 5: NBA Shot Locations

With X and Y coordinates it is possible to plot the different zones on a basketball chart to identify patterns and understand the different zones better.

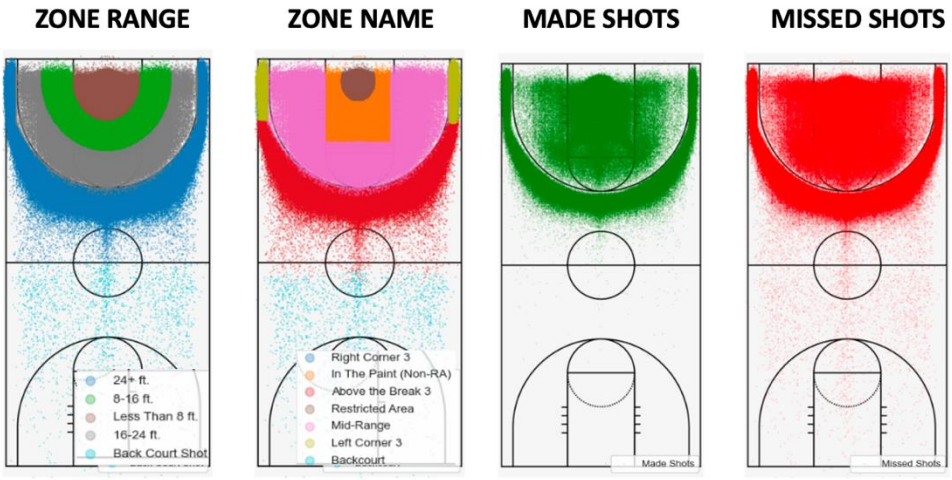

Figure 6: Geospatial Analysis of Shots

A geospatial analysis can also be done by studying the relation of coordinates to shot success. As can be portrayed in Figure 6, there is a high density of shots (both made and missed) close to the basket, which is typical because shots are generally easier to make from a shorter distance. The red dots are spread throughout the court, showing that players miss from all areas, but there seems to be a higher concentration of missed shots from the mid-range to the 3-point line. The green dots

also appear throughout but are less densely packed as you move away from the basket, which is consistent with the lower shooting percentages generally seen with longer shots. Overall, these shot charts suggest a modern basketball approach that emphasizes efficiency, with a focus on three-pointers and shots at the rim, which are generally considered the most efficient shots in basketball.

Shot Distance Distribution

The histogram in Figure 7 shows that shot distances are heavily concentrated at the lower end, with a steep decline as distance increases. This is expected as most shots in basketball are taken closer to the basket. The box plot does not show any extreme outliers, which suggests that most shot distances are within a reasonable range for basketball shots. However, there is a long tail towards the higher distances, indicating that long shots are less frequent but still within the realm of normal play.

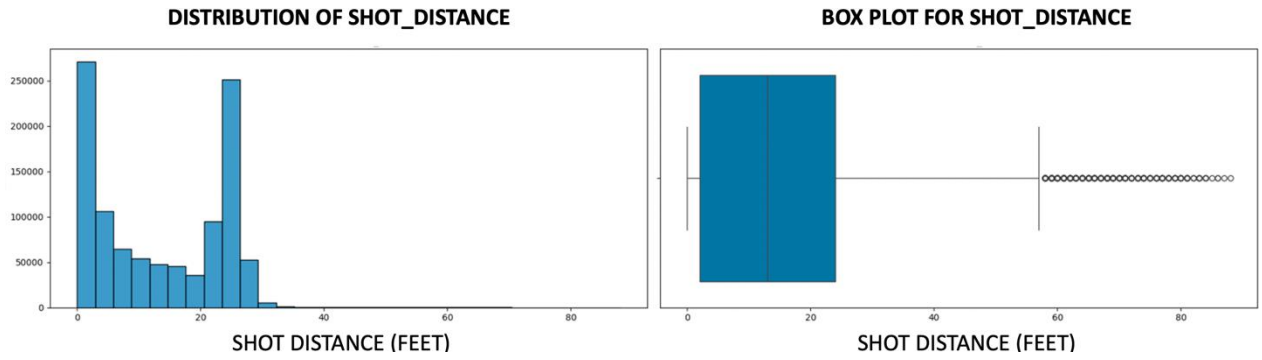

Figure 7: Distribution of Shot Distance

The mean shot distance is approximately 13.5 feet, with a standard deviation of about 10.6 feet. This large standard deviation indicates a wide variety of shot distances. The median shot distance is 13 feet, which is slightly less than the mean, suggesting a slight skew in the distribution toward longer shots. The maximum shot distance recorded is 88 feet, which is an exceptionally long shot, likely attempted at the end of a quarter or game.

Statistics from Numerical Features

The time-related features do not show any clear trend or pattern that would indicate a direct impact on the likelihood of a shot being made. However, it's possible that in combination with other features these time variables could still provide valuable information when used in a predictive model.

Table 1 shows a summary and description of the numerical features in the NBA dataset using quantitative measures. These statistics are explored with the purpose of providing and overview of the dataset through metrics such as mean, median, standard deviation, minimum, maximum, quartiles and counts. The summary helps us view the central tendency, dispersion and basic properties of each variable.

	QUARTER	MINS_LEFT	SECS_LEFT	LOC_X	LOC_Y	SHOT_DISTANCE
count	1.032499e+06	1.032499e+06	1.032499e+06	1.032499e+06	1.032499e+06	1.032499e+06
mean	2.481068e+00	5.368144e+00	2.887049e+01	4.383952e-02	1.009674e+01	1.346666e+01
std	1.133216e+00	3.456007e+00	1.742016e+01	7.247337e+00	7.317195e+00	1.058196e+01
min	1.000000e+00	0.000000e+00	0.000000e+00	-2.500000e+01	5.000000e-02	0.000000e+00
25%	1.000000e+00	2.000000e+00	1.400000e+01	-1.200000e+00	5.965000e+00	2.000000e+00
50%	2.000000e+00	5.000000e+00	2.900000e+01	0.000000e+00	7.015000e+00	1.300000e+01
75%	3.000000e+00	8.000000e+00	4.400000e+01	1.100000e+00	8.550000e+00	2.400000e+01
max	8.000000e+00	1.200000e+01	5.900000e+01	2.490000e+01	9.045000e+01	8.800000e+01

Table 1: Summary Statistics of Numerical Features

3.3 Exploratory Data Analysis (EDA)

An Exploratory Data Analysis was crucial for our research and understanding of the dataset prior to building the shot prediction models. This analysis focused on visually summarizing and understanding the main characteristics of the NBA dataset regarding distribution, outlier detection, variable relationships, patterns or anomalies.

Studying the correlation between the different variables within the dataset can help explore which values have more impact and importance for determining the success of a basketball shot. SHOT_DISTANCE has the strongest negative correlation with SHOT_MADE (− 0.225). This

indicates that as shot distance increases, the likelihood of making the shot decreases, which aligns with general basketball understanding. NORM_LOC_Y also has a negative correlation (− 0.182), suggesting that shots taken further from the basket in the y-axis are less likely to be successful. Temporal variables like MINS_LEFT, and SECS_LEFT have a very weak positive correlation with SHOT_MADE, indicating that shots taken later in the game or quarter might be slightly more likely to be successful. Other variables have very low correlation values, suggesting that they might not have a strong linear relationship with shot success. However, they could still be predictive in conjunction with other features or might need to be transformed or combined with other variables to reveal their predictive power.

A detailed examination through violin plots reveals a clear relationship between shot distance and success rates (Figure 8). The plots show a higher density of successful shots at shorter distances, a logical outcome in basketball, where proximity to the basket typically increases the likelihood of scoring. This trend diminishes with increasing distance, aligning with the inherent challenges of long-range shooting. However, a notable deviation from this trend appears around the 23-feet mark, as indicated by a higher density of successful shots at this distance. This anomaly can be attributed to the three-point line in basketball. The three-point line, typically set at around 23 feet from the basket in professional leagues, offers players the opportunity to score three points with a single shot, as opposed to the standard two points for shots taken within the line. This rule incentivizes players to develop proficiency in shooting from this specific range, hence the increased density of successful shots at around 20 feet.

Figure 8: Shot Distance VS Shot Made

The bar plot analysis of shots made by game quarter reveals consistent success rates across the first four quarters, with minor variations (Figure 9). However, there's a noticeable decrease in success rates during overtime periods, possibly due to factors like player fatigue, heightened defensive pressure, or selective shooting. The data from overtime periods, particularly the third and fourth, should be interpreted cautiously due to smaller sample sizes which could skew the results.

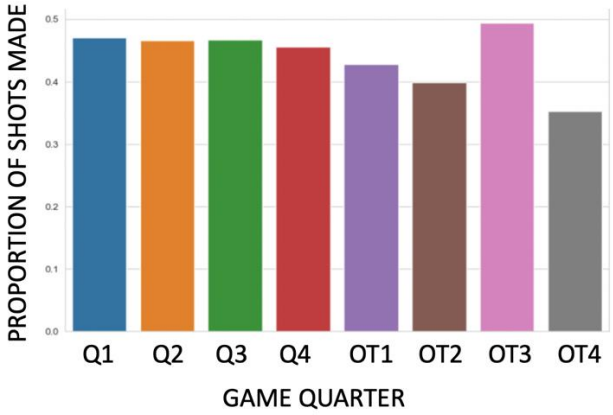

Figure 9: Shot Success Rate by Game Quarter

Additionally, the time left in the quarter when a shot is taken plays a significant role in success rates. Shots in the final moments of a quarter often have lower success rates, likely due to the hurried nature of these attempts. As the time left increases, so does the success rate, stabilizing for shots taken well before the quarter's end. Interestingly, shots taken at the very beginning of a quarter show a slight dip in success rate, hinting at less strategic, more spontaneous shot selections.

Regarding the relationship between action type and shot success, it is established that the action types with the highest success rates are predominantly various types of dunk shots, which are known to have a high likelihood of scoring. Moreover, as can be seen in the bar chart in Figure 10, the shot success rate is highest in the restricted area, significantly lower for mid-range and paint shots, and further reduced for three-point shots, with the lowest success rate being for backcourt shots.

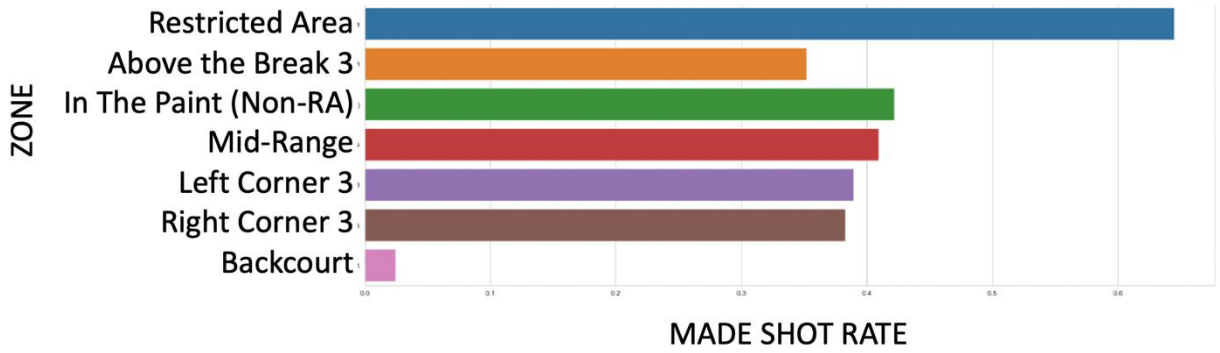

Figure 10: Shot Success Rate by Zone

A first estimation of the relationship between SHOT_MADE and the numerical features could easily be established with the correlation matrix. It showed that the attributes with the highest correlation to focal variable SHOT_MADE are LOC_Y, MINS_LEFT, QUARTER, and SECS_LEFT with respective correlations to shot made 0.11, 0.02, 0.01, and 0.01.

A key finding is the relationship between shot distance and success rate, with a higher success rate for shots closer to the basket, except for a notable increase at the three-point line. The data also reveals a consistent shot success rate across game quarters, with a decline during overtime, potentially due to player fatigue. Additionally, the attributes LOC_Y, MINS_LEFT, QUARTER, and SECS_LEFT show the highest, albeit low, correlations with the shot made variable, indicating their potential predictive value.

3.4 Data Cleaning

Prior to fitting the models, it is necessary that the data is cleaned and preprocessed to ensure accuracy and reliability in the subsequent analyses. This section contains a general overview of the preprocessing steps that apply to all models discussed in this paper. A more in-depth overview of data preprocessing as well as additional preprocessing steps are discussed for each model separately in the respective methodology sections.

Data Cleaning

Data cleaning is a crucial preliminary step in preparing the dataset for analysis. It ensures the accuracy and reliability of the data, which is essential for the effectiveness of the models developed from this data. The NBA shots dataset underwent a thorough cleaning process, addressing several key issues to enhance its quality and usability.

Handling Missing Values

An initial assessment of the dataset revealed missing values in the POSITION_GROUP and POSITION columns, accounting for approximately 0.51% of the data. Given the relatively small proportion of missing data and its negligible impact on the dataset's overall integrity, the decision was made to remove these rows. This approach simplifies the dataset and avoids the potential biases or inaccuracies that might arise from imputation or other methods of dealing with missing data.

Elimination of Duplicate Records

Duplicate records can skew data analysis, leading to erroneous conclusions. In this dataset, a total of 60 instances of duplicate records were identified. To preserve the integrity of the dataset and ensure that each entry represents a unique shot attempt, these duplicates were removed. This step is vital for maintaining the dataset's validity, ensuring that subsequent analyses are based on accurate and unique data points.

Treatment of Outliers

Outliers in the LOC_X, LOC_Y, and SHOT_DISTANCE columns were another significant concern. Specifically, 45,957 outliers were found in the 'LOC_X' column, 8,780 in the 'LOC_Y' column, and 1,749 in the 'SHOT_DISTANCE' column. However, upon closer examination, it was determined that these outliers fall within the expected range of a basketball court. They represent valid, albeit less common, long-range shot attempts. Recognizing the importance of these data points for a realistic representation of shot distribution, the decision was made to retain these outliers. This inclusion ensures that the dataset comprehensively covers the full spectrum of shot types, including those that are statistically rare but strategically significant.

Streamlining the Dataset

In addition to addressing missing values, duplicates, and outliers, the data cleaning process also involved streamlining the dataset by removing certain attributes. This step was taken to focus the dataset on the most relevant variables and to simplify the analysis. The attributes shown in Table 2 were removed given that they either provide redundant information, are not directly relevant to shot analysis, or can be better represented in a transformed or aggregated form.

Attribute	Justification
SEASON_2	The same as season 1 (but with different formatting)
TEAM_NAME	Attribute TEAM_ID contains the same information
PLAYER_NAME	Attribute PLAYER_ID contains the same information
GAME_DATE	Broken down into SEASON, MONTH, DAY, WEEKDAY
GAME_ID	No predictive power (only used in LSTM model to create sequences)
HOME_TEAM and AWAY_TEAM	Integrated into one binary column IS_HOME_TEAM (where 1 = home, 0 = away)
EVENT_TYPE	Equal to SHOT_MADE/SHOT_MISSED
ZONE_NAME	Zone name is already captured in ZONE_ABB
MINS_LEFT	Integrated with SECS_LEFT into feature SECS_LEFT
SECS_LEFT	Updated with MINS_LEFT column

Table 2: Deleted Features

Hence, the data cleaning process for the NBA shots dataset was thorough and considered various aspects crucial for maintaining data quality. By addressing missing values, duplicates, and outliers,

and by streamlining the dataset through the removal of certain variables, the dataset is now primed for more detailed feature engineering and subsequent analysis.

3.5 Data Preprocessing

The quality of predictions hinges on the transformation and feature engineering of the available data. In this context, data transformation encompasses converting diverse data types, such as numerical and date-time variables, into a consistent format. Meanwhile, feature engineering involves the creation of new informative variables that capture nuanced insights from the raw data. These preprocessing techniques not only enhance the interpretability of the data but also empower ML algorithms to extract meaningful patterns.

In the initial phase of preprocessing, categorical identifiers such as team IDs, player IDs, and zone abbreviations were converted into numerical formats. This conversion is essential for ML algorithms, which typically require numerical input. A significant transformation was the creation of the binary IS_HOME_TEAM column, indicating whether the team is playing at home. This not only simplifies the feature but also reduces memory usage, making future processing more efficient.

Additionally, the SHOT_MADE variable was converted to an integer format (1 for true, 0 for false), and SHOT_TYPE was transformed into a binary variable. Numerical continuous variables underwent scaling to normalize their ranges, which is particularly important for algorithms sensitive to variable scales. Encoding of categorical variables was a crucial aspect of this phase. One-hot encoding was applied to variables like TEAM_ID, PLAYER_ID, POSITION, POSITION_GROUP, ACTION_TYPE, BASIC_ZONE, and ZONE_ABB. This process transformed these variables into a binary matrix representation, essential for numerical analysis in ML models. Meanwhile, ordinal encoding was utilized for variables such as SEASON and

QUARTER to preserve any inherent order in these categories. The temporal aspect of the data was also addressed. The GAME_DATE column was converted into a datetime object, allowing for the extraction of additional features like the day of the week and the month of the game. Meanwhile, the SEASON_1 column, which provided year information, was retained and renamed as SEASON.

A key transformation was made to the time-related variables. The TIME_LEFT column was computed by converting the minutes left (MINS_LEFT) into seconds and adding the remaining seconds (SECS_LEFT). This unified representation of time as a single numeric value in seconds is more intuitive and conducive to analyzing time-related patterns in the data. Spatial variables, particularly LOC_X and LOC_Y, showed different scaling across years. To address this and ensure comparability across different seasons, these variables were normalized to the scale of a basketball court.

The optimization of memory and data type conversion was also undertaken. Numerical columns were cast to more memory-efficient data types, with Float64 columns converted to float32 and int64 columns to int32. Categorical columns were explicitly converted to categorical types, optimizing memory usage and improving the performance of ML models. The final integration step involved dropping the original categorical columns from the cleaned DataFrame and incorporating the newly encoded features. Through these comprehensive transformations and optimizations, the NBA shots dataset was effectively prepared for ML analysis, ensuring that the data is not only accurate and relevant but also structured in a way that maximizes the potential insights derived from subsequent modeling. Table 3 shows the final features used after data cleaning and processing.

Feature	Description
SEASON	Categorical variable indicating season
TEAM_ID	Categorical variable indicating NBA team
PLAYER_ID	Categorical variable indicating individual player
POSITION_GROUP	Categorical variable indicating the position of the player who shoots (e.g. F)
POSITION	Categorical variable representing a more detailed description of the position of the shooting player (e.g. SF)
SHOT_MADE	Numerical (binary) variable indicating whether shot was made (successfully = 1)
ACTION_TYPE	Categorical variable indicating the type of action performed with an attempted shot such as three pointing, layup or hook shot
SHOT_TYPE	Numerical (binary) variable indicating the type of shot (I.e. three-point or two-point)
BASIC_ZONE	Categorical variable indicating the zone on the field from which is shot (e.g. Right Corner 3 or Above the Break 3)
ZONE_ABB	Categorical variable indicating the specific zone on the zone (e.g. right, right center)
ZONE_RANGE	Categorical variable indicating the distance to the hoop (e.g. 8-16 ft, 24+ ft.)
SHOT_DISTANCE	Numerical variable indicating the distance ot the hoop
QUARTER	Categorical variable indicating the particular quarter of the game in which a shot is made
IS_HOME_TEAM	Numerical (binary) variable indicating whether the team of the attempted shot is the home team
MONTH	Categorical variable indicating the month (ranging from 1 to 12)
DAY	Numerical variable indicating the day
WEEKDAY	Categorical variable indicating the day of the week (ranging from 1 to 7)
TIME_LEFT	Numerical variable representing total time left in seconds (relative to the quarter)
NORM_LOC_X	Numerical (float) variable indicating the normalized x-coordinate relative to the size of the basketball field
NORM_LOC_Y	Numerical (float) variable indicating the normalized y-coordinate relative to the size of the basketball field
TOTAL_TIME_LEFT	Numerical variable representing total time left in seconds (relative to the whole game)

Table 3: Final Features

3.6 Data Limitations

The NBA shots dataset from 2019 to 2023, while a rich resource for predicting shot success, has inherent limitations and potential biases that impact the interpretability and applicability of predictive models developed from it. One significant limitation is the absence of visual spatial data, such as that provided by SportVU systems. The dataset includes basic two-dimensional spatial data (LOC_X, LOC_Y) but lacks the more nuanced three-dimensional spatial context that visual spatial

data offers. This advanced data captures the dynamics of player movements, spacing, and defensive positioning, crucial elements that influence shot success.

Another key limitation is the lack of information regarding the game score at the time of each shot. Understanding whether a team is leading or trailing, and by how many points, is vital as it can significantly influence a player's decision-making and shot success. Players might behave differently under various game pressures, such as taking riskier shots when trailing or opting for safer options when leading. This game context is a critical factor that, if included, could provide deeper insights into shooting effectiveness under varying game conditions.

The dataset's temporal scope, covering seasons from 2019 to 2023, also poses limitations. While it provides a substantial amount of recent data, this range may not capture the long-term trends and evolutions in playing styles, strategies, and player skills. The nature of basketball evolves over time, and patterns from earlier years might offer valuable insights that are not present in a dataset limited to recent seasons. Furthermore, the dataset lacks deeper contextual information about the players and teams involved, despite including PLAYER_ID and TEAM_ID. Factors such as player fatigue, recent performance, injury status, team dynamics, and coaching strategies can all significantly impact a player's shot success. For instance, a player's shooting efficiency might decrease towards the end of a tough series of games due to fatigue, a factor not directly captured in the dataset. There is also a potential bias in the representation of players and teams and does not consider external factors like crowd presence, referee decisions, and environmental conditions for outdoor venues. These factors, though challenging to quantify, can subtly influence game dynamics and player performance.

In summary, while the NBA shots dataset provides valuable data for shot success prediction, its limitations necessitate a cautious approach in the interpretation of analysis or predictive modeling

results. Addressing these limitations through the inclusion of additional data sources and considering them in model development could lead to more accurate and robust predictions.

3.7 Data Management and Reproducibility

In the realm of Business Analytics, the dataset utilized for our research holds academic relevance and is well-structured, making it an ideal candidate for comprehensive analysis. This dataset, primarily focused on NBA shots, was obtained from a GitHub repository maintained by Dominic Samangy, a Basketball Analytics Coordinator for the New Orleans Pelicans. The repository, a treasure trove of information, not only provides access to the dataset but also includes additional resources that are instrumental in aiding data exploration and visualization. It's important to note that the original source of this dataset is the NBA.com website, which is the official platform for NBA statistics. This origin ensures the authenticity and reliability of the data, crucial for any analytical endeavor. The management, storage, and accessibility of the data are key components in ensuring reproducibility of our research. To maintain efficient version control and manage our coding processes GitHub was utilized.

As we ventured into more intensive computational tasks, especially after the preprocessing phase, we encountered the challenge of handling significantly large dataframes. This was particularly evident after the encoding of categorical variables, a process that notably increased the size and complexity of the data. To address this challenge, we leveraged the computational power provided by the Server of Nova Data Science Knowledge Lab (DSKC). The DSKC server was crucial in managing these heavy processes, such as fitting complex models, which required substantial computational resources beyond what typical personal computing environments offer.

Access to the DSKC server was facilitated through remote connection software and SSH (Secure Shell), providing a flexible and secure means of utilizing these resources. This access not only

allowed us to handle large datasets efficiently but also ensured that the data processing and model fitting were conducted in a robust and reproducible manner. The combination of GitHub for version control and the DSKC server for computational power forms a comprehensive data management strategy, essential for the integrity and reproducibility of our research. This approach not only guarantees the accuracy of our current analysis but also sets a solid foundation for any future studies that may build upon our work.

4. Methods

This section of our thesis dives into the advanced models built to forecast NBA shot prediction. Four different ML algorithms, two ensemble models (Random Forests and XGBoost) and two neural networks (Neural Network Long Short-Term Memory and Recurrent Neural Network) are the foundation of our methodology. All four models are highly appropriate for our end goal as they bring different viewpoints to the field of predictive analytics. After thoroughly examining every model, results, accuracy, performance, strengths, and limitations of each model are discussed.

4.1 Evaluation Metrics

Prior to exploring the different models for predicting the success of shots made in the NBA, the metrics are determined by which the models will be assessed. In predictive modeling, especially in sports analytics, selecting the right evaluation metrics is crucial to assess the effectiveness of the models accurately. While accuracy is the primary metric due to its widespread use in existing literature (Appendix A), we also consider ROC AUC as a metric and have a look at the confusion matrix to capture different aspects of model performance.

- **Accuracy**: Accuracy measures the proportion of correct predictions (both true positives and true negatives) out of all predictions made. It is calculated as *(True Positives + True*

Negatives) / Total Predictions (Meehan 2022). In the context of NBA shot prediction, accuracy tells how often the model correctly predicts whether a shot is successful or not. A high accuracy indicates that the model is effective in distinguishing between made and missed shots. It is the primary metric because it offers a straightforward and intuitive understanding of model performance and allows for easy comparison with other studies in basketball analytics (Appendix A).

- **ROC AUC**: The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) curve is a performance measurement that assesses the ability of a model to distinguish between classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. In shot prediction, the AUC-ROC score helps understand how well the model differentiates between made and missed shots (Shah and Romijnders 2016).
- **Confusion matrix**: The confusion matrix is a tabular representation that categorizes the predictions of a model into four types: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). This matrix is particularly valuable in providing a clear and concise visual representation of the model performance (Meehan 2022).

By employing these metrics, we aim to provide a thorough and nuanced assessment of our predictive models. This multi-faceted approach ensures a robust evaluation of our NBA shot prediction models.

GROUP PART

4.2 Neural Networks

This chapter builds upon the findings presented in Chapter 2.4, where a thorough review of studies highlighted the potential of neural networks in improving the performance of basketball shot prediction. In this section, a comprehensive overview of neural networks is provided, and the data preprocessing steps necessary to use neural networks for analytical purposes is detailed. The chapter then focuses on the application of two types of neural networks: Feedforward Neural Networks (FNNs) and Recurrent Neural Networks (RNNs), which were identified in the literature review before (Chapter 2.4). Through this exploration, we seek to not only validate the capabilities of these networks but also to contribute to the ongoing evolution of sports analytics methodologies.

Neural networks are computing systems inspired by biological neural networks. These systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. Neural networks consist of layers of interconnected nodes, with each node representing a neuron and each connection representing a synapse (Dongare et al. 2012). When the connection value is high, it suggests that a strong connection is given. The strength of these connections, or weights, is adjusted during training, allowing the network to learn from the input data (Prieto et al. 2016). In the context of NBA shot prediction, neural networks can process and analyze vast amounts of data, such as player positions and historical shot success rates, to predict the likelihood of a shot being successful. There are various reasons to use neural networks in classification problems, the most relevant points are:

- **Handling of High-Dimensional Data**: Their ability to process complex, multi-factorial data makes them ideal for applications like NBA shot prediction (Prieto et al. 2016).

- **Feature Extraction and Learning**: Neural networks automatically identify relevant features from data, essential for accurate classification in sports analytics. (Zhang 2000).
- **Flexibility and Adaptability**: They can be tailored to various data types and tasks, adding versatility to predictive modeling (Zhang 2000).
- **Ability to Model Non-linear Relationships:** The complexity of data often involves non-linear relationships, which neural networks can effectively model (Zhang 2000).

Neural networks come in various architectures, each suited to specific tasks or problems. The most important types in basketball shot prediction that were identified in Chapter 2.4 include:

- **Feedforward Neural Networks (FNNs):** The simplest type of artificial neural network architecture, where connections between the nodes do not form a cycle (Meehan 2017; Harmon et al. 2021).
- **Recurrent Neural Networks (RNNs)**: RNNs have connections that form cycles, allowing information to persist and are useful for processing sequential data (Shah and Romijnders 2016; Yan et al. 2023; Wang & Zemel 2016).

In the extensive literature review presented in Chapter 2.4, we explored the potential of neural networks in enhancing the accuracy of shot prediction in basketball, which highlighted various research gaps. Moving on, we will primarily explore the application of FNNs and RNNs, chosen for their relevance and proven effectiveness in shot prediction. Our goal is to provide structured methodologies for addressing the research gaps identified, contributing to the advancement of sports analytics, particularly in the nuanced area of shot prediction.

4.2.1 Uniform Data Preprocessing for Neural Networks

In the preparation of data for neural network models, several critical preprocessing steps were undertaken to ensure the integrity and uniformity of the input data. These steps are fundamental for both Feedforward Neural Networks (FNNs) and Recurrent Neural Networks (RNNs) used in our shot prediction analysis.

Firstly, all numerical features underwent z-score normalization to standardize their distribution. This ensures that each feature has a mean of zero and a standard deviation of one, facilitating the gradient-based optimization algorithms that are central to neural network training. Such standardization is imperative, especially for features with different scales (for instance, the different scaling of X_LOC/Y_LOC compared to SECS_LEFT), allowing for a smoother and more efficient gradient descent process.

For categorical variables, we utilized one-hot encoding to transform categories within variables like TEAM_ID, PLAYER_ID, POSITION, and POSITION_GROUP into separate binary features. This approach prevents the neural network from misinterpreting the ordinal significance of numerical categories and ensures that each entity is represented distinctly, enhancing pattern recognition. We also computed the TOTAL_TIME_LEFT for each shot, which is especially crucial for RNNs. This calculation provides the model with a temporal context, crucial for understanding the game's flow and the strategic shifts that occur, particularly in the closing stages of the game or overtime periods.

Normalization of spatial variables LOC_X and LOC_Y to a consistent basketball court scale was applied, allowing the networks to interpret these features uniformly across different seasons. This step is vital for our models to generalize well across various playing conditions. Additionally, the transformation of categorical variables such as SEASON and QUARTER into ordinal categories

and their subsequent encoding maintains the chronological order of the data. This is particularly significant for LSTM models within the RNN family, which can then discern patterns and trends across different seasons and stages of the game.

In summary, our comprehensive preprocessing approach, tailored to the nuances of neural networks, ensures that the data fed into our models is optimized for pattern recognition and predictive accuracy. Whether it is the spatial and temporal aspects for FNNs or the sequential and time-sensitive nature for RNNs, these preprocessing steps form the foundation upon which our neural network models are built and trained, setting the stage for robust and insightful shot prediction analysis in NBA games.

INDIVIDUAL CONTRIBUTION OF SEBASTIAN MANI VARADAPPA

4.2.2 Feedforward Neural Networks

4.2.2.1 Introduction to FNN

Feedforward Neural Networks (FNN) stand as a cornerstone in neural network architecture. Their orderly, linear configuration of layers makes them highly adaptable for various applications, including predictive analytics in sports like basketball (Ivankovic et al. 2010).

An FNN is architecturally composed of several layers: an input layer, one or more hidden layers, and an output layer. Each layer consists of neurons, which are interconnected across successive layers (Hansson & Olsson 2017). In a basketball context, the input layer handles raw data inputs such as player positions, shot distances, time remaining or game location (home or away). The hidden layers are where FNNs perform their computational feats. Neurons within these layers apply weights and biases to the inputs and pass them through an activation function. Activation functions like ReLU (Rectified Linear Unit) or Sigmoid are pivotal, introducing non-linearity into the network. This non-linearity allows the network to discern complex patterns in the data. The

Sigmoid function, often used for binary classification, outputs values between 0 and 1, making it suitable for scenarios like predicting the probability of a basketball shot's success (Szandała 2020). ReLU, known for its efficient training performance, activates neurons only when the input is positive, helping in faster convergence (Dudek 2021). Other activation functions, such as Tanh or Softmax, have their unique characteristics and are chosen based on the specific needs of the model (Goodfellow et al. 2016).

The number of hidden layers and neurons within each layer are critical hyperparameters, influencing the network's capacity to learn and make accurate predictions. Data flows in a forward direction through the network, from input to output, with the final layer tailored for the specific task at hand. For predicting basketball shot success, the output layer would generally consist of a single neuron with a Sigmoid function, yielding a probability of success (Wang et al., 1992).

Training an FNN is a process of refining the weights and biases to minimize the disparity between predicted and actual outcomes. This is typically achieved through backpropagation and optimization algorithms like Stochastic Gradient Descent (SGD) or Adaptive Moment Estimation (Adam). Backpropagation efficiently calculates the loss function's gradients - a critical measure of prediction error - relative to the weights, enabling the network to adjust its weights to reduce the error (Kurilovich, 2022).

In terms of evaluating FNN performance, metrics like log loss (or cross-entropy loss) are commonly used, particularly for binary classification tasks. Log loss measures the dissimilarity between the actual labels and the predicted probabilities, offering a robust way to gauge the model's accuracy. A lower log loss value indicates better performance, with the model's predictions closely aligning with the actual data (Qin et al., 2019).

FNNs excel in processing and learning from large datasets, identifying intricate patterns that might escape human analysts. In basketball, a FNN can analyze extensive historical shot data, considering a wide range of variables influencing shot success. This enables the FNN to detect subtle patterns, like variations in a player's success rate based on factors such as positioning, opposition, venue, or time in the game. This depth of analysis underpins the power of FNNs in sports analytics and beyond. However, it's important to note that while FNNs are robust in pattern recognition, they have limitations in processing sequential or time-series data. Unlike Recurrent Neural Networks (RNNs), FNNs cannot capture temporal dependencies as they treat each input independently. This limitation is significant in applications like sports analytics, where understanding the sequence of events is essential (Goodfellow et al. 2016).

4.2.2.2 Methodology for FNN

Our process commenced with a strategic split of data into training, testing, and validation sets to mitigate potential bias and validate the model's accuracy. Following this, we proceeded with meticulous data preprocessing and feature selection, which is vital for the network's learning efficiency. Once the data was prepared and the relevant features identified, we turned our focus to the model's architecture. We carefully determined the optimal configuration of layers and neurons to effectively capture the dynamics influencing shot success. Lastly, the evaluation of the FNN's performance was based on accuracy and AUC scores, providing a comprehensive assessment of its predictive accuracy in real-game scenarios.

4.2.2.3 Data Preprocessing for FNN

Building on the uniform data preprocessing steps outlined in section 4.3.1, additional preprocessing was undertaken specifically for the Feedforward Neural Network (FNN) model to enhance its

predictive accuracy. A key distinction in our approach was the introduction of embedding layers for Player IDs and Team IDs.

The decision to use embeddings in the FNN model was influenced by the network's ability to interpret and utilize dense representations of categorical data. Embedding layers transform player and team identifiers, which are high-cardinality categorical variables, into meaningful, lower-dimensional vectors. This is particularly beneficial for FNNs as it enables the model to discern intricate patterns and influences of individual players and teams on shot success. By efficiently handling these variables, embeddings aid in capturing complex relationships, leading to better model generalization and predictive performance, without the computational burden of one-hot encoding (Goodfellow et al. 2016).

The specific inclusion of embeddings in the FNN model, as opposed to the RNN model, aligns with each network's architectural strengths. While RNNs excel in processing sequential data, FNNs are better suited to discerning patterns from aggregated, non-temporal data. Thus, embeddings provide a more efficient way to input high-cardinality categorical data into the FNN, optimizing its performance (LeCun et al. 2015).

4.2.2.4 Feature selection for FNN

Considering the shortcomings identified in Meehan's (2017) study, our feature selection process is curated to include both categorical and numerical data. This inclusion ensures that our model benefits from the rich, diverse inputs that reflect the multidimensional nature of the game.

Initially, the dataset contained 131 features post-preprocessing. To determine their relative importance, we employed a Random Forest Classifier, which is effective for feature ranking due to its ensemble learning approach. This method offers an intuitive understanding of feature

significance, as it evaluates the predictive power of each feature across numerous decision trees (Ardeti et al. 2022).

Through this analysis, we identified the most influential features, which guided our initial model testing. We experimented with models using varying numbers of top-ranked features to assess performance. However, it was observed that incorporating a higher number of features consistently yielded better results. This led to the realization that in the context of neural networks, which inherently weigh features during training, including all features was the most effective strategy. The network's ability to assign appropriate weights minimized the impact of less significant features, thereby optimizing model performance without necessitating the exclusion of potentially useful data.

Our decision to use the entire set of 131 features is supported by the neural network's architecture, which is adept at managing and interpreting a large number of inputs. This approach ensures that the model can leverage the full spectrum of data, allowing for a more nuanced and accurate prediction of NBA shot success.

4.2.2.5 Model Architecture for FNN

Our model architecture is deliberately designed to capture the complexity inherent in the task of shot prediction. The foundation of our approach lies in the division of data into distinct sets for training, validation, and testing. We employ an 80/10/10 split, allocating 80% of the data to training, with the remaining 20% equally divided between validation and testing. This distribution ensures a thorough learning process while providing a rigorous assessment of the model's predictive capabilities.

The architecture of our FNN begins with separate input layers for player and team data, leveraging **embedding layers** to effectively handle the high cardinality of these categorical variables. These

embeddings are then flattened and concatenated with the other features from the dataset, creating a rich input for the subsequent neural network layers. The first **hidden layer** of our network consists of 64 neurons using the **ReLU** activation function for its ability to maintain gradient flow during training. Following this, we incorporated an additional **dense layer** with 64 neurons to add complexity to the model. Both hidden layers are equipped with a **dropout rate of 40%**, serving as a regularization technique to prevent overfitting. The final output layer comprises a single neuron employing a **sigmoid** activation function, offering a probability output that indicates the likelihood of a shot being successful.

To compile our model, we employed the **Adam** optimizer with a **learning rate of 0.0001**, chosen for its effectiveness in adjusting weights during training. The **binary cross-entropy** loss function was used for both training and validation phases, aligning with our binary classification objective. Furthermore, the inclusion of an early stopping mechanism, with a patience of **20 epochs**, fortifies our model against overfitting by stopping training when the validation loss ceases to decrease, ensuring our model's generalizability.

After the training is complete, we evaluate our FNN using the test data, focusing on log loss and accuracy metrics. These metrics are crucial as they not only measure the model's predictive performance but also the confidence in its probability estimates, ensuring that our model predictions are both precise and reliable.

By integrating these methodological elements, we aim to construct a robust and reliable neural network model. This model is not only grounded in a thorough analysis of NBA shot data over several seasons but is also built upon an architecture and evaluation framework designed to address the nuanced challenges of predicting complex events such as basketball shots.

4.2.2.6 Results for FNN

This section outlines the performance of the Feedforward Neural Network (FNN) developed for NBA shot prediction. The model's effectiveness is evaluated using the primary dataset from seasons 2019-2023 and a secondary dataset for the 2015 season, providing insights into both current and historical predictive capabilities.

Model Performance on Test Set (Seasons 2019-2023)

We assessed the FNN using the test set, which represents 20% of the data. This evaluation provided a robust platform to gauge the model's effectiveness.

		Predicted	
		Negative	Positive
Actual	Negative	19784	9129
	Positive	28172	45629

Figure 11: FNN Confusion Matrix for Test Set 2019-2023

The confusion matrix (Figure 11) shows the classification ability of the model. The FNN achieved a test accuracy of 63.68% and a test loss of 0.6342. Additionally, the AUC-ROC score for this test period was 0.6662, reflecting the model's ability to effectively differentiate between successful and unsuccessful shot attempts. These results are promising, as they suggest the model can reliably predict shot outcomes in a range of game situations. The test accuracy and AUC score indicate a robust level of predictive performance, aligning well with the complexities of in-game shot prediction scenarios.

Model Performance on Test Set (Season 2015)

The model was further validated using data from the 2015 NBA season. In this historical test, the FNN showed a test accuracy of 62.77% and a test loss of 0.6460, with an AUC score of 0.6603.

		Predicted	
		Negative	Positive
Actual	Negative	27379	11619
	Positive	64855	101583

Figure 12: FNN Confusion Matrix for 2015 Season

The consistency of the model's performance across different seasons underscores its versatility and adaptability. This ability to maintain prediction accuracy over time suggests the model's algorithms and feature selection are robust against the variances in play styles and strategies that evolve across seasons.

In summary, the FNN model exhibits solid predictive performance, as evidenced by its accuracy and AUC scores in both the recent and historical datasets. The results not only affirm the model's reliability but also highlight its potential applicability in diverse basketball analytics scenarios. The consistency of its performance, despite the evolving nature of the game, positions the FNN as a valuable tool for teams and analysts looking to leverage data-driven insights for strategic decision-making.

4.2.2.7 Discussion for FNN

Our FNN model, implemented for both the 2015 season and the 2019-2023 seasons, demonstrated promising results, achieving a test accuracy of 63.68% and an AUC score of 0.6662 for the 2019-2023 dataset, and a slightly lower test accuracy of 62.77% and an AUC score of 0.6603 for the 2015 dataset. These results are particularly significant when compared to previous literature in the field of NBA shot prediction using FNNs.

When we compare our model's performance with that of previous studies, such as Meehan's (2017) research which achieved 55%, we observe a notable improvement. Meehan's study was limited by the exclusion of categorical data and a lack of spatial detail. In contrast, our model incorporated

these categorical variables, which contributed to its enhanced performance. This suggests that the inclusion of rich, context-specific data, especially categorical variables, plays a crucial role in improving the predictive accuracy of FNN models in sports analytics.

An interesting finding from our research was that expanding the dataset to include multiple seasons did not significantly enhance the model's performance. A similar level of accuracy was achieved when the model was trained solely on the 2019 season's data. This indicates that the quantity of data, in terms of spanning multiple seasons, may not be as critical as the quality and relevance of the data used.

Despite the advances our FNN model represents over previous studies, there remains a noticeable gap when compared to our baseline models. This gap underscores the potential for further enhancements in our approach to NBA shot prediction. One key area of potential improvement lies in the quality of the game-specific data used. The inclusion of more nuanced variables such as defender distance, shot clock time, and possibly other in-game contextual factors could provide a richer dataset for the model to learn from. Such detailed data could capture the subtleties and complexities inherent in basketball shots more effectively.

In addition to data enrichment, there is also an opportunity to explore alternative data preprocessing methods. Our comprehensive efforts in hyperparameter tuning and feature selection have led us to conclude that potential improvements in our model's performance may be less about the model's internal settings and more about the nature and preprocessing of the input data. Although we have already explored several preprocessing methods like one-hot encoding and embedding techniques for high cardinality categorical data, there could be alternative approaches or more advanced techniques that have the potential to yield better performance.

Moving forward, we recommend focusing on acquiring richer, more detailed game-specific data and experimenting with innovative data preprocessing methods. Adding nuanced variables such as defender distance and shot clock time could significantly enhance the model's understanding of shot contexts. These areas present promising avenues for future researchers to explore, potentially leading to more refined predictive models in the domain of basketball shot prediction.

GROUP PART

5. Discussion

This chapter discusses the performances of the predictive models, emphasizing the critical role of feature selection and the unique strengths of each model type. The chapter also explores the substantial business impacts and research contributions stemming from improved prediction accuracy. Additionally, the chapter acknowledges the limitations of the study, such as data constraints and computational power, and suggests directions for future research, offering a comprehensive overview of the models' efficacy in sports analytics.

5.1 Interpretation of Results

The evaluation of four distinct models for NBA shot prediction - Random Forest, XGBoost, Feedforward Neural Network, and Recurrent Neural Network with LSTM layer – has produced encouraging results. A comparative analysis of these models offers valuable insights into their theoretical and practical applications. Notably, the four central models in this paper were enhanced by increasing the volume of data points used, the inclusion of more categorical features and more extensive hyperparameter optimization. An overview of the performance results for all models can be found in Table 4.

Model	**Accuracy**	**ROC – AUC**
Random Forest	63.19 %	0.6689
XGBoost	63.60 %	0.6111
FNN	62.77 %	0.6603
LSTM	70.88 %	0.7734

Table 4: Summary Test Results on 2015 Dataset

Both ensemble models, Random Forests and XGBoost, exhibited effective predictive abilities, achieving accuracies in the range of 63-64%. These findings align with the established strengths of ensemble methods, including their robustness and capability to handle complex data structures. Notably, our study reveals that expanding the dataset to encompass multiple seasons did not significantly impact the performance of these models, indicating the primacy of data quality over sheer quantity.

Moreover, the introduction of more categorical variables positively influenced the predictive power of these models, as evidenced by their enhanced accuracy and feature importance rankings. This suggests a re-evaluation of traditional metrics in sports analytics, incorporating these novel elements for a more comprehensive analysis.

Additionally, our extensive hyperparameter optimization process led to only marginal improvements, highlighting the nuanced balance between model complexity and its effective tuning. The XGBoost model, in particular, validated the importance of spatial and categorical features such as player position, action type, and shot zone range as critical for accurate shot prediction, resonating with findings from existing studies. Similarly, the Random Forest model underscored the significant yet often overlooked influence of categorical factors like 'DAY' and 'WEEKDAY' on basketball performance, suggesting potential new avenues for sports analytics research. These results reinforce the value of incorporating a rich variety of categorical data in predictive models for a deeper and more nuanced understanding of game dynamics.

Regarding the DL models, the FNN and RNN models demonstrated notable accuracies of 63% and 71% respectively. These results represent significant improvements over previous studies, such as models by Meehan (2017) and Oughali et al. (2020). Notably, the inclusion of a broader range of data across multiple seasons did not significantly impact the FNN model's performance. However, the RNN with LSTM showed remarkable improvement, indicating that for certain types of neural network models, particularly those handling sequential data, the scope of data can be a critical factor in enhancing predictive accuracy. This distinction is further evidenced by the superior performance of our LSTM model over the FNN, suggesting that while FNNs are adept at processing large datasets, they exhibit limitations in handling sequential data, a key aspect of sports analytics.

In contrast, RNNs with LSTM layers excel in managing sequential data, instrumental in identifying complex temporal patterns in sports events. This proficiency in handling sequential data is a significant reason for the superior performance of LSTM over FNN, and even the two ensemble models, in our analysis. The LSTM model's success, with its distinct advantage in processing time-

sensitive data, underscores the untapped potential of RNNs in basketball analytics, demonstrating their superior accuracy and adaptability.

The positive influence of incorporating more categorical data was evident in both FNN and LSTM models, as seen in their enhanced accuracy and feature importance rankings. This highlights the critical role of rich, context-specific data in improving the predictive accuracy of neural network models in sports analytics. However, despite extensive hyperparameter tuning and optimization, improvements in the FNN and LSTM models were only marginal. This suggests that, alongside data quality and the inclusion of categorical variables, there is a nuanced balance to be struck in model architecture and parameter optimization to fully realize the potential of these sophisticated neural networks.

5.2 Business Implications

As already mentioned in the introduction of our thesis (Chapter 1), an improvement of the existing basketball shot prediction models by even 1% can already be very advantageous for many NBA teams. However, with the introduction of a LSTM model we were able to achieve an improvement of 2%-points, which could potentially lead to an even bigger impact on a team's finances.

To underline this, we looked at the case of the Phoenix Suns. For the 2023-24 season, the Suns have a midrange shot success rate of 47.2% (NBA Media Ventures). We assume that 2%-points improvement in prediction accuracy leads to a 2% increase in shot success rate. This assumption is based on the premise that enhanced predictive accuracy directly translates to more effective training and in-game strategies. This could involve adjusting shooting techniques, optimizing player positioning, and tailoring training regimens to reinforce the most effective shooting patterns. Additionally, improved prediction models can offer insights into opponent defenses, enabling teams to exploit weaknesses more effectively. These results in a new success rate of approximately

48.144%. This seemingly small increase of 0.944% can have a significant impact on the team's performance and, by extension, its financial health. With the Suns attempting an average of 85.4 field goals per game (NBA Media Ventures), a 0.944% increase in shot success translates to 0.806 additional successful shots per game. If these shots are evenly distributed between 2 and 3-pointers, this could mean an additional 2.015 points per game on average. In the highly competitive landscape of the NBA, even such marginal gains are crucial. For instance, 5.2% of the games in the current season have been decided by just 1 point. For a team like the Phoenix Suns, playing 82 games per season, these enhancements in shooting accuracy could potentially turn close games in their favor. With 5.2% of games being close enough to be affected by this improvement equates to about 4 additional wins per season. In the context of the NBA, where the competition is fierce, these additional wins could be the difference between making or missing the playoffs or securing a more favorable seeding.

From a financial perspective, based on the findings from the paper by Li (2011), each additional win can bring about a 0.3% increase in team revenue. For the Suns, with a revenue of $516 million in 2023 (Ozanian 2023), this could mean an additional $6.19 million (4 wins x 0.3% of $516 million) in revenue. This increase is not just from ticket sales but also from the higher marketability and increased fan engagement that typically accompany a winning team. It's important to note that these calculations are based on a specific set of assumptions and only calculated for the case of the Phoenix Suns. The actual impact could vary based on numerous factors, including team dynamics, the strength of opponents, and changes in player performance. Nonetheless, this analysis underscores the potential financial benefits of even slight improvements in prediction model, highlighting the intricate link between on-court success and off-court financial gains.

Beyond the direct financial implications for NBA teams like the Phoenix Suns, there are other business impacts of improving basketball shot prediction models by 2%. Advancements in predictive modeling can have implications for sports betting markets, where accuracy in predicting game outcomes is highly valued. Improved models can offer more reliable data for bookmakers and bettors alike, potentially influencing betting patterns and market dynamics. In the realm of player development and scouting, better prediction models can aid teams in identifying talent more effectively and make more informed decisions during drafts, trades, and in developing training programs tailored to enhance specific skills of players.

5.3 Limitations of the Study

While our study has provided valuable insights into basketball shot prediction, it is crucial to acknowledge certain limitations that may impact the interpretation of our findings and the generalizability of our results. Our study primarily relies on player tracking data, which may differ from studies incorporating visual data sources such as SportVU. Previous research leveraging visual data has demonstrated the potential for achieving higher accuracies in shot prediction. Therefore, it is important to consider that variations in data sources may contribute to differences in model performance and outcomes.

Another limitation stems from computational power constraints. Our analysis covers 2019-2023 seasons, but with enhanced computational resources, we could potentially extend our investigation to include all 20 seasons from 2003 to 2023. The restricted timeframe may influence the comprehensiveness of our findings, and future research with greater computational capabilities could explore a broader temporal scope. Notably, certain important factors that have been considered in other analyses, such as dribbles, final margin, shot clock time, and closest defender distance, were not included in our dataset. The absence of these variables may impact the depth

and granularity of our analysis, potentially limiting the scope of our insights in comparison to studies that incorporate a more extensive set of features.

These limitations underscore the need for future research to address these constraints, potentially incorporating diverse data sources, leveraging extended timeframes, and including a broader array of relevant variables. Despite these limitations, our study contributes valuable findings to the existing body of knowledge on basketball shot prediction and serves as a foundation for further exploration in this dynamic field.

6. Conclusion

This thesis has successfully demonstrated the efficacy of four ML techniques including XGBoost, Random Forest, Feedforward Neural Network (FNN) and Recurrent Neural Network (RNN) with an LSTM layer in predicting NBA shot outcomes, achieving improvements over existing models. In addition, the four models have each established the significance of certain features in predicting shot success; the top three features with the most predictive power being shot distance, action type, x and y coordinates of the player on the field and the time left.

Addressing the limitations of existing studies predicting shot success, our thesis explored the effects of the volume of data points used, the breadth of feature selection, and hyperparameter optimization of models. Our analysis revealed that while hyperparameter tuning only marginally improved model performances, the incorporation of more categorical data had a significantly positive impact on our results. Interestingly, expanding the dataset size did not uniformly enhance model performance; it notably benefited the LSTM model, but had limited impact on others. This suggests that isolating individual players and teams, or other patterns that span multiple seasons, is not effectively captured by including data from five seasons, or such factors might be insignificant in predicting shot success in the NBA.

The significant breakthrough of this research lies in the exploration and validation of the Recurrent Neural Network (RNN) model, particularly adept in predicting NBA shot success due to their ability to process sequential data. RNNs excel in capturing both temporal and spatial dynamics, crucial in basketball where the sequence of movements leading up to a shot is often predictive of its outcome. Specifically, we have seen that RNNs that incorporate Long Short-Term Memory (LSTM) units, adeptly manage the challenges of conventional RNNs in learning long-term dependencies. This is vital for comprehending the complex play sequences in basketball. The integration of LSTM units enables the RNN model to not only capture the immediate context of a shot but also the broader play patterns, leading to a top-1 classification accuracy of 71% in our study. Notably, in the complex context of basketball, where numerous unpredictable elements can influence a player's shooting performance, expecting significantly high accuracies (around 90%) with the available features may be unrealistic.

This advancement is especially relevant in the business sphere of the NBA league, where financial performance is increasingly reliant on the precision of data-driven strategies. We have estimated that our improved model with a marginal increase of 2 percentage points in performance, relative to existing NBA models, could translate into substantial monetary gains, primarily through improved shooting success rates. We showcased this in our case study about the Phoenix Suns with a gain of $6.19M in the current season. More accurate predictions enable coaches to tailor training and strategies more effectively, leading to higher shot accuracy. This improvement in player performance can lead to increased fan engagement, higher ticket sales, merchandising opportunities, and lucrative sponsorships.

Building on the strengths of our study, which already benefits from an increased data size compared to previous research, future investigations should aim to further exploit this advantage by extending

the dataset to encompass a larger number of seasons. This expansion would not only provide a richer pool of data points for enhanced predictive accuracy but also allow for increased predictive power of player-specific variables. Another promising area for future research involves integrating more dynamic spatial features, especially from datasets like SportVU, into the RNN model specifically. RNNs are particularly adept at processing and interpreting complex spatial relationships such as detailed player movements and court positions. Consequently, by feeding more comprehensive spatial data into RNNs, researchers could significantly enhance the RNN's capability to understand and predict shot outcomes in the context of the dynamic spatial environment of a basketball game.

In summary, this thesis underscores the transformative impact of advanced predictive models in the realm of sports analytics. By demonstrating the effectiveness of ML and deep learning techniques, particularly the potential of the RNN, this research contributes to the ongoing evolution of data-driven decision-making in sports, with far-reaching implications for team performance, fan engagement, and financial success in the NBA league.

7. References

Akers, Michael D., Shaheen Wolff and Thomas E. Buttross. 1992. "An empirical examination of the factors affecting the success of NCAA Division I college basketball teams." *Accounting Faculty Research and Publications*, 72.

Alanazi, Hamdan O., Abdul H. Abdullah and Kashif N. Qureshi. 2017. "A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care." *Journal of Medical Systems*, 41(4), 69. https://doi.org/10.1007/s10916-017-0715-7

Alhnaity, Bashar and Maysam Abbod. 2020. "A new hybrid financial time series prediction model." *Engineering Applications of Artificial Intelligence*, 95, 103873. https://doi.org/10.1016/j.engappai.2020.103873

Ardeti, V. A., Kolluru, V. R., Varghese, G. T., & Patjoshi, R. K. 2022. "An Outlier Detection and Feature Ranking based Ensemble Learning for ECG Analysis." International Journal of Advanced Computer Science and Applications (IJACSA). https://thesai.org/Publications/ViewPaper?Volume=13&Issue=6&Code=IJACSA&SerialNo=86

Beal, Ryan, Timothy J. Norman and Sarvapali D. Ramchurn. 2019. "Artificial intelligence for team sports: A survey." *The Knowledge Engineering Review*, 34. https://doi.org/10.1017/s0269888919000225.

Bradlow, Eric T., Manish Gangwar, Praveen Kopalle and Sudhir Voleti. 2017. "The role of big data and predictive analytics in retailing." *Journal of Retailing*, 93(1), 79–95. https://doi.org/10.1016/j.jretai.2016.12.004

Breiman, Leo. 1996. "Bagging predictors." *Machine Learning*, 24(2), 123-140. https://doi.org/10.1007/BF00058655

Breiman, Leo. 2001. "Random forests." *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bunker, Rory and Teo Susnjak. 2019. "The application of machine learning techniques for predicting results in team sport: A review." *Journal of Artificial Intelligence Research*, 73, 1285–1322. https://doi.org/10.1613/jair.1.13509

Chen, Tianqi and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Conferences*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Tianqi and Tong He. 2023. "xgboost: eXtreme Gradient Boosting." *The Comprehensive R Archive Network*. https://cran.ms.unimelb.edu.au/web/packages/xgboost/vignettes/xgboost.pdf

Chen, Wei-Sen, and Yin-Kuan Du. 2009. “Using neural networks and data mining techniques for the financial distress prediction model.” *Expert Systems with Applications*, 36(2), 4075–4086. https://doi.org/10.1016/j.eswa.2008.03.020

Chmait, Nader, and Hans Westerbeek. 2021. “Artificial intelligence and machine learning in sport research: An introduction for non-data scientists.” *Frontiers in Sports and Active Living,* 3. https://doi.org/10.3389/fspor.2021.682287

Cranmer, Skyler J. and Bruce A. Desmarais. 2017. “What can we learn from predictive modeling?” *Political Analysis*, 25(2), 145–66. https://doi.org/10.1017/pan.2016.5

Dongare, A.D., R.R. Kharde, and A.D. Kachare. 2012. “Introduction to artificial neural network.” *International Journal of Engineering and Innovative Technology (IJEIT)*, 2(1), 189–194.

Dudek, G. 2021. “Data-Driven Learning of Feedforward Neural Networks with Different Activation Functions.” Springer. https://link.springer.com/chapter/10.1007/978-3-030-87986-0_6

Elkan, Charles. 2010. “Predictive analytics and data mining.” San Diego: University of California.

Freund, Yoav, and Robert E Schapire. 1997. “A decision-theoretic generalization of on-line learning and an application to boosting.” *Journal of Computer and System Sciences*, 55(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

Fregoso-Aparicio, Luis, Julieta Noguez, Luis Montesinos, and José A. García-García. 2021. “Machine learning and deep learning predictive models for type 2 diabetes: a systematic review.” *Diabetology & Metabolic Syndrome*, 13, 148. https://doi.org/10.1186/s13098-021-00716-w

García, Javier, Sergio J. Ibáñez, Raúl Martinez De Santos, Nuno Leite, and Jaime Sampaio. 2013. “Identifying basketball performance indicators in regular season and playoff games.” *Journal of Human Kinetics*, 36, 161–168. https://doi.org/10.2478/hukin-2013-0015

Geisser, S. 1993. “Predictive inference: An introduction.” New York: Chapman & Hall.

Giri, Chandadevi, Ulf Johansson and Tuwe Löfström. 2019. “Predictive modeling of campaigns to quantify performance in fashion retail industry.” *2019 IEEE International Conference on Big Data (Big Data)*, 2267–73. https://doi.org/10.1109/BigData47090.2019.9005918

Goodfellow, Ian, Yoshua Bengio and Aaron Courville. 2016. “Deep Learning.” MIT Press.

Harmon, Mark, Abdolghani Ebrahimi, Patrick Lucey, and Diego Klabjan. 2021. “Predicting shot making in basketball learnt from adversarial multiagent trajectories.” *International Journal of Sport and Health Sciences,* 11.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. “Long short-term memory.” *Neural Computation*, 9(8), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

Huyghe, Thomas, Pedro E. Alcaraz, Julio Calleja-González, and Stephen P. Bird. 2022. “The underpinning factors of NBA game-play performance: A systematic review (2001-2020).” *The Physician and Sportsmedicine*, 50, 94–122. https://doi.org/10.1080/00913847.2021.1979182

Ivankovic, Z., Rackovic, M., Markoski, B., Radosav, D., & Ivkovic, M. 2010. “Analysis of basketball games using neural networks.” Proceedings of the 11th International Conference on Computer Information Systems and Industrial Management Applications (CINTI). https://ieeexplore.ieee.org/document/5672237

Krebs, J. 2022. “Rapsodo, Driveline, and the potential of analytics implementation into professional sports.” Retrieved from https://static1.squarespace.com/static/62058eb2a4c56b1dd98f8c0c/t/6239fb2ef952311bc62100cc/1647967022999/Rapsodo%2C+Driveline%2C+and+the+Potential+of+Analytics+Implementation+into+Professional+Sports.pdf

Kuhn, M., and K. Johnson. 2019. “Applied predictive modeling.” New York: Springer.

Kumar, V., and M.L. Garg. 2018. “Predictive analytics: A review of trends and techniques.” *International Journal of Computer Applications*, 182(1), 31–37. https://doi.org/10.5120/ijca2018918212

Kurilovich, P. 2022. “Using population algorithms to optimize the objective function when training artificial neural networks.” IEEE. https://ieeexplore.ieee.org/document/9873490

Labarère, José, Renaud Bertrand, and Michael J. Fine. 2014. “How to derive and validate clinical prediction models for use in intensive care medicine.” *Intensive Care Medicine*, 40, 513–527. https://doi.org/10.1007/s00134-014-3212-3

Li, Hongfei, and Maolin Zhang. 2021. “Artificial Intelligence and Neural Network-Based Shooting Accuracy Prediction Analysis in Basketball.” *Mobile Information Systems.* https://doi.org/10.1155/2021/4485589

Li, Harrison. 2011. "True Value in the NBA: An Analysis of On-Court Performance and Its Effects on Revenues." Department of Economics, University of California, Berkeley.

Li, Y., and Feng, T. 2020. “The effects of sport expertise and shot results on basketball players’ action anticipation.” PLOS ONE, 15(1), e0227521. https://doi.org/10.1371/journal.pone.0227521

Liu, Q., Cui, X., Abbod, M.F., Huang, S.-J., Han, Y.-Y., and Shieh, J.-S. 2011. “Brain death prediction based on ensembled artificial neural networks in neurosurgical intensive care unit.” *Journal of Taiwan Institute of Chemical Engineer*s, 42, 97–107. https://doi.org/10.1016/j.jtice.2010.03.020

Maglott, J. C., D. Chiasson and P.B. Shull. 2019. “Influence of skill level on predicting the success of one's own basketball free throws.” *PLOS ONE*, 14(3), e0214074. https://doi.org/10.1371/journal.pone.0214074

Meehan, Brett. 2017. “Predicting NBA shots.” Stanford. https://cs229.stanford.edu/proj2017/final-reports/5132133.pdf

Minusha, S. R. 2016. “Sports analytics.” Summit Research Repository. https://summit.sfu.ca/item/16939

Nakai, Y., et al. 2023. “Video-based basketball shooting prediction and pose correction using machine learning.” Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-14490-2

NBA Media Ventures. “NBA Advanced Stats”. Accessed December 13, 2023. https://www.nba.com/stats/teams/shots-general?PerMode=PerGame&Season=2023-24

Oughali, Maram S., Mariah Bahloul and Sahar A. El Rahman. 2019. “Analysis of NBA Players and Shot Prediction Using Random Forest and XGBoost Models.” [Conference Proceedings]. https://dx.doi.org/10.1109/ICCISCI.2019.8716412

Ozanian, Mika and Justin Teitelbaum. „NBA Valuations” Forbes. October 26, 2023. https://www.forbes.com/lists/nba-valuations/

Prieto, Alberto, Beatriz Prieto, Eva Martinez Ortigosa, Eduardo Ros, Francisco Pelayo, Julio Ortega, and Ignacio Rojas. 2016. “Neural Networks: An Overview of Early Research, Current Frameworks and New Challenges.” *Neurocomputing*, 214, 242–268. https://doi.org/10.1016/j.neucom.2016.06.014

Qin, Z., Zhang, Z., Li, Y., & Guo, J. 2019. “Making Deep Neural Networks Robust to Label Noise: Cross-Training With a Novel Loss Function.” IEEE Access. https://ieeexplore.ieee.org/document/8834773

Rein, R. and D. Memmert. 2016. “Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science.” *SpringerPlus* 5, 1410. https://doi.org/10.1186/s40064-016-3108-2

Scikit-learn. “sklearn.ensemble.RandomForestClassifier.” Accessed December 06, 2023. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Shah, Rajiv and Rob Romijnders. 2016. “Applying Deep Learning to Basketball Trajectories.”

Sherstinsky, Alex. 2020. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network.” *Physica D: Nonlinear Phenomena*, 404, 132306. https://doi.org/10.1016/j.physd.2019.132306

Skinner, Brian. 2012. “The Problem of Shot Selection in Basketball.” Edited by Matjaz Perc. *PLoS ONE*, 7(1), e30776. https://doi.org/10.1371/journal.pone.0030776

Smith, M. and R. Singh. 2023. “Advancing predicted feedback for improved motor training. Harvard Office of Technology Development.” https://otd.harvard.edu/explore-innovation/technologies/advancing-predicted-feedback-for-improved-motor-training

Steinberg, Leigh. 2015. “Changing the game: The rise of sports analytics.” Forbes. https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/

Szandała, T. 2020. “Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks.” Springer. https://link.springer.com/book/10.1007/978-981-15-5495-7

TensorFlow. “tf.keras.utils.pad_sequences.” September 27, 2023. https://www.tensorflow.org/api_docs/python/tf/keras/utils/pad_sequences

TensorFlow. “Keras: The high-level API for TensorFlow.” June 08, 2023. https://www.tensorflow.org/guide/keras

Thabtah, Fadi, Li Zhang, and Neda Abdelhamid. 2019. “NBA Game Result Prediction Using Feature Analysis and Machine Learning.” *Annals of Data Science*, 6, no. 1: 103–16. https://doi.org/10.1007/s40745-018-00189-x.

Vicent, Jose F., Enrique Moreno, and David Gil. 2022. “Is the Future of Basketball Being Influenced by Predictive Data Analysis?” *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4308292

Wang, Kuan-Chieh and Richard Zemel. 2016. “Classifying NBA Offensive Plays Using Neural Network.” MIT Sloan Sports Analytics Conference. https://www.cs.toronto.edu/~zemel/documents/1536-Classifying-NBA-Offensive-Plays-Using-Neural-Networks.pdf

Wang, Z., Tham, M., & Morris, A. 1992. “Multilayer feedforward neural networks: a canonical form approximation of nonlinearity.” International Journal of Control. https://www.tandfonline.com/doi/abs/10.1080/00207179208934333

Wright, R., J. Silva and I. Kaynar-Kabul. 2016. “Shot recommender system for NBA coaches.”

Yan, Wenlin, Xianxin Jiang, and Ping Liu. 2023. “A Review of Basketball Shooting Analysis Based on Artificial Intelligence.” *IEEE Access*, 11: 87344–87365. https://doi.org/10.1109/ACCESS.2023.3304631

Zhang, G. P. 2000. “Neural Networks for Classification: A Survey.” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 30, 451-462. https://doi.org/10.1109/5326.897072

Zhang, Shaoliang, Alberto Lorenzo, Changjing Zhou, Yixiong Cui, Bruno Gonçalves, and Miguel Angel Gómez. 2019. “Performance profiles and opposition interaction during game-play in elite basketball: Evidences from National Basketball Association.” *Int. J. Perform. Anal. Sport*, 19, 28–48. https://doi.org/10.1080/24748668.2018.1555738.

8. List of Figures

9. List of Tables

10. Appendix

Appendix A: Summary of Existing Literature on Predicting Shot Success

Paper	Data Source	Data Size	Type of Model	Features used	Performance metric(s)	Performance Value(s)
Predicting NBA Shots by Meehan (2022)	NBA 2014-2015	42.105	Logistic regression	dribbles, period, final margin, touch time, shot clock, closest defender distance, and shot distance	Accuracy	59%
	NBA 2014-2015	42.105	Random Forest	dribbles, period, final margin, touch time, shot clock, closest defender distance, and shot distance	Accuracy	61%
	NBA 2014-2015	42.105	XGBoost	dribbles, period, final margin, touch time, shot clock, closest defender distance, and shot distance	Accuracy	68%
	NBA 2014-2015	42.105	SVM	dribbles, period, final margin, touch time, shot clock, closest defender distance, and shot distance	Accuracy	55%
	NBA 2014-2015	122.502	FNN Neural Network (sigmoid only)	dribbles, period, final margin, touch time, shot clock, closest defender distance, and shot distance	Accuracy	55%
	NBA 2014-2015	122.502	FNN Neural Network (sigmoid and RELU)	dribbles, period, final margin, touch time, shot clock, closest defender distance, and shot distance	Accuracy	55%
Shot Recommender System for NBA Coaches by Wright et al. (2016)	NBA 2015-2016	Not given	Factorization Machine	Not given	RMSE	Not given

Analysis of NBA Players and Shot Prediction Using Random Forest and XGBoost Models by Oughali et al. (2019)	NBA 2014-2015	203.591	Random Forest	the shot clock, the number of dribbles, the shot distance, and the closest defender's distance	Accuracy	57%
	NBA 2014-2015	203.591	XGBoost	Final margin, period, shot clock, number of dribbles, shot distance, touch time, closest defender's distance, PTS Type, FGM, PTS	Accuracy	68%
Applying Deep Learning to Basketball Trajectories by Shah and Romijnders (2016)	SportVU NBA data 2015-2016	20.000	XGBoost	X (length of court), Y (width of court), Z (height of ball), game clock variables, difference in movement over each time period for each dimension, distance to the center point of the rim, difference over time for this distance, and the angle of the ball with respect to the rim.	AUC	0.719
	SportVU NBA data 2015-2016	20.000	RNN Neural Network	X (length of court), Y (width of court), Z (height of ball), game clock variables, difference in movement over each time period for each dimension, distance to the center point of the rim, difference over time for this distance, and the angle of the ball with respect to the rim.	AUC	0.843
Predicting shot making in basketball learnt from adversarial multiagent trajectories by Harmon et al., (2016)	SportVU NBA 2012-2013	75000	FNN Neural network	Player and ball positions at the time of the shot, Game time and quarter time left on the clock, Player speeds over five seconds, Speed of the ball, Distances and angles (with respect to the hoop) between players,	Error Rate, Log Loss	0.72, 0.81

Predicting shot making in basketball learnt from adversarial multiagent trajectories by Harmon et al., (2016)	SportVU NBA 2012-2013	75000	FNN Neural network	Number of defenders in front of the shooter (300 angle of the shooter) and within six feet based upon the angles calculated between players, Ball possession time for each offensive player, Number of all individuals near the shooter (including teammates)	Error Rate, Log Loss	0.72, 0.81
Classifying NBA offensive plays using neural networks by Wang & Zemel (2016)	SportVU NBA 2013-2014	Not given	RNN Neural Network	Players' coordinates on the court, 3D position of the ball at 25 frames per second, unique player identifier.	Top 1 Accuracy, Top 3 Accuracy*	0.66, 0.806

*The first metric is top-1 accuracy, which compares the single highest-scoring class by the model to the correct answer, on each test example. Top-3 accuracy considers the k classes that attain the highest scores for the model on that example.

Appendix B: NBA Dataset Data Types and Description Before Processing

Variable Name	Description	Sample	Data Type		
TEAM_ID	Unique ID per team	1610612762	Integer	Nominal	Categorical
PLAYER_ID	Unique ID per player	1628960			
GAME_ID	Unique ID per game	21801229			
POSITION	Player position	SG	Object		
POSITION_GROUP	Player position group	G			
ACTION_TYPE	Description of shot type (layup, dunk, jump shot, etc.)	Jump Shot			
SHOT_TYPE	Type of shot (2PT or 3PT)	3PT Field Goal			
ZONE_ABB	Abbreviation of the side of court	R			
BASIC_ZONE	Name of the court zone the shot took place in	Right Corner 3			
ZONE_NAME	Name of the side of court the shot took place in	Right Side			
ZONE_RANGE	Distance range of shot by zones	24+ ft.			
TEAM_NAME	Name of Team	LA Clippers			
PLAYER_NAME	Name of Player	Tyrone Wallace			
HOME_TEAM	Game played at team's venue	LAC			
AWAY_TEAM	Visiting Team	UTA			
EVENT_TYPE	Character variable denoting a shot outcome	Made Shot			
GAME_DATE	Date of game (M-D-Y)	04-10-2019			
SHOT_MADE	X coordinate of the shot in the x, y plane of the court (0, 50)	True	Boolean		
QUARTER	Quarter of the game	1	Integer	Ordinal	
SEASON_1	Season indicator	2019			
SEASON_2	Season indicator	2018-19	Object		
LOC_Y	Y coordinate of the shot in the x, y plane of the court (0, 50)	12.15	Float	Numerical	
LOC_X	Distance of the shot with respect to the center of the hoop, in feet	- 22.2			
SHOT_DISTANCE	Seconds remaining in minute of the quarter	23	Integer		
SECS_LEFT	Minutes remaining in the quarter	46			
MINS_LEFT	Date of game (M-D-Y)	1			

Appendix C: GitHub Repository

https://github.com/monica-navas-m/NBA_ShotPrediction