Predicting the 2024 NBA Championship With Machine Learning [Team 27]

Zareb Islam, Aadhil Mubarak Syed, Jaynor Singson, Dylan Tran
Department of Computer Science
University of California, Davis, Davis CA, USA 95616
Email: {zaislam, amubaraksyed, jasingson, dyltran} @ucdavis.edu

Abstract*—This paper presents the development and evaluation of a predictive model for forecasting the NBA Championship of the 2024 season. This predictive model will use deep neural networks, a common Machine Learning technique. Using a dataset consisting of historical and current team performance metrics, game outcomes, and advanced analytics, our model aims to determine the likelihood of a specific NBA team winning the championship for the 2024 NBA season. We employed a combination of multiple machine learning algorithms, including linear and logistic regression, in our exploratory data analysis to identify the most significant predictors of championship success. Our model performance was evaluated using cross-validation methods and achieved an exceptional accuracy rate. Additionally, we performed comparative analysis of different methods to find the most effective model. The results indicated that our prediction model can serve as a valuable tool for analysts and fans alike in the basketball community, providing insights into the statistics and metrics that could potentially influence championship outcomes. Future work will focus on refining the model by incorporating real-time data and expanding the feature set to enhance predictive accuracy, as well as for future championship predictions in upcoming seasons.

I. Introduction and Background

The game of basketball is dynamic and unpredictable, with continuous evolution in team strategies, player performance, and game statistics. It is indubitable that players as a whole possess a lot more talent today compared to older eras. For instance, in the recent 2024 season, the Golden State Warriors held the 10th seed in the Western Conference despite having 45 wins. To put this into perspective, 45 wins would have awarded them the 4th seed in the previous year's standings. With such talent in the league, traditional methods of predicting season outcomes, based on subjective analysis or simple statistical approaches, often fall short in accuracy. A prime example of this can be found during last year's postseason when the Miami Heat, an 8th seed in the Eastern Conference, defeated many teams that were title favorites on their path to the NBA finals, taking the entire NBA fanbase by surprise. Given such stochasticity, it is tempting to conclude that statistics do not tell the entire story of a team's outcome... or do they?

The recent surge in Machine Learning (ML) offers a sophisticated way to handle vast amounts of data and uncover patterns that may not be immediately apparent, providing a more robust method for prediction. In this project, we aim to use Machine Learning to predict which team will be crowned the 2024 NBA Champions. By employing various ML techniques and algorithms, we aim to analyze various features of overall and advanced team performance statistics from historical data to make informed predictions for the outcomes of the current season.

In this project, our objective is to develop a predictive model that is able to accurately predict the 2023/2024 season's NBA Championship winner. To accomplish this, we will utilize a multitude of historical and real-time data to provide a strong predictive framework. For our learning algorithm, we contemplated various techniques, including multiple linear regression, Bayesian probability networks, and deep neural networks (DNN). DNNs, capable of handling complex patterns in large datasets, are suitable for the nuanced analysis of sports statistics. On the other hand, regression models, while simpler, can provide quick insights into relationships between variables and outcomes. After thorough analysis of the advantages and disadvantages of each of the algorithms, we decided on using DNNs as our learning algorithm.

DNNs are a subset of artificial intelligence and machine learning. Neural network architectures are capable of automatically extracting and learning features from otherwise nonlinear data through multiple layers of abstraction, enhancing the predictive power. Each layer captures a different feature. From simple edges in early layers to the more complex patterns in deeper layers. This form of hierarchical learning allows the model to recognize the complex relationships and structures in large sets of data. In the context of the NBA, where data richness and variability are prominent, DNNs can provide deeper insights and more accurate predictions compared to traditional models.

The NBA holds a rich history of data, which includes decades of historical records, statistics, accomplishments, and performance metrics. Some key statistics include team metrics like win-loss records and offensive/defensive ratings. Some player statistics include points per game, assists, and rebounds. Advanced analytics include the likes of effective field goal percentage and true shooting percentages. This kind of data can be accessed through different websites and online resources. Our project uses Basketball Reference (https://www.basketball-reference.com/) as our source of data.

Basketball analysts use these advanced analytics and metrics to assess how strong certain teams are. These statistics are key to predicting winners of individual games as well as predicting the overall champion of the NBA. The problem we aim to address is building a strong model to predict the

NBA Championship with high confidence. As important as statistics are, our goal is to find the key statistics and metrics in evaluating possible championship winners. By doing so, we aim to create a predictive framework that is both accurate and insightful.

Predicting the NBA Championship winner presents a difficult task. The NBA is an ever-changing landscape, and predicting the winner for a singular season can be deemed almost impossible due to unpredictable factors such as injuries, mid-season trades, and suspensions as well. By using historical data and continuously learning from new data, we can use deep neural networks to adapt to these trends that may not be easily quantifiable. This adaptive learning process is important to maintain model accuracy throughout the NBA season.

This project seeks to highlight the critical factors that lead to championship success and showcase the effectiveness of deep learning in sports prediction. Our results could offer valuable insights to analysts, coaches, and teams, promoting a data-driven approach to strategic planning and performance enhancement. Furthermore, this project contributes to the broader field of sports analytics by demonstrating the application of advanced machine learning techniques for complex challenges, such as predictive analytics. By expanding our knowledge in the application of data-driven models in sports, we hope to pave the way for future innovations and applications in this dynamic field.

II. Literature Review

Predicting outcomes in professional competitions and sports such as the NBA playoffs has remained an important topic in the field of machine learning. This literature review explores the past developments in the application of machine learning to NBA outcomes, with the intention to compare to our own project.

The availability of datasets from sources like Basketball Reference has been key to fueling detailed analyses and model training. Yang and Lu (2012) used many standard team statistics as features, including averages in points, turnovers, field goals, and rebounds. Houde (2021) found that plus/minus, offensive rating, defensive rating, true shooting percentage, and win percentage over the past ten games were important features in predicting NBA game outcomes. He also utilized an elo rating that was calculated after every game in order to approximate team strength. Allen (2023) constructed a target variable called “champion share” that measured team success, counting the number of playoff wins each team achieved and dividing that value by the maximum sixteen possible wins. Features that he found to be extremely important in predicting championship teams included being a top three seed, offensive and defensive ratings, presence of an MVP player, shooting statistics, performance against strong teams, and playoff experience.

The decision of what machine learning model to implement is also critical to prediction success. Among the models implemented include Logistic Regression, Linear Discriminant Analysis, Random Forest, Classification and Regression Tree, Gaussian Naive Bayes Network, Support Vector Machines, K-Nearest Neighbors, and XGBoost Classifier. Studies often employ metrics such as accuracy, precision, recall, and F1 score to gauge the effectiveness of predictive models. Additionally, cross-validation techniques are commonly used to assess models. Liu (2021) found the Linear Discriminant Analysis model performing the best and Houde (2021) found the Gaussian Naive Bayes model performing best.

Building upon these advancements, our project leverages a deep neural network model to predict NBA championship winners. This network feeds upon a manually coded target values dataset indicating championship outcomes, something we created that is unique to our project. Possible values include 0 (did not make playoffs, or lost in conference quarterfinals), 0.125 (conference semifinals), 0.25 (conference finals), 0.5 (NBA finals), and 1.0 (NBA champion) to indicate how far each team made it on their respective runs towards the championship.

III. Dataset Description and Exploratory Data Analysis

The datasets for this project are sourced from Basketball-Reference (https://www.basketball-reference.com/), a website containing “statistics, scores, and history for the NBA, ABA, WNBA, and top European competition.” We made use of the per-game team statistics and the advanced team statistics for every season. Because the game has changed dramatically since its beginning in the mid-1900s, we decided that using data from this time would likely make our model less accurate. Therefore, we will be using datasets spanning specifically the last 30 years (from 1993 to 2023).

Initially, we considered employing web scraping techniques to collect the data; however, after encountering unexpected complexities with this approach, we opted to manually extract the data by copying CSV data directly from the website. During pre-processing, we eliminated and combined columns that were irrelevant or could cause collinearity in our predictive model, such as rank, minutes played, 3P%, FT%, total rebounds, wins, and losses. To differentiate team data from opponent data, we prefixed opponent metrics with an ”O” (e.g., ”PPG” became ”OPPG”). We merged the datasets using the team name as the key and included a year column to distinguish team entries across seasons.

In our exploratory data analysis (EDA), we focused on comparing the statistical profiles of NBA championship-winning teams to league averages across various seasons through a series of 20 pre-determined questions. This comprehensive examination allowed us to identify key trends influencing a team’s likelihood of winning a championship.

One of the most notable trends is the evolution in style of play, embodied by an increased pace and a higher rate of three-point attempts among championship teams over time. Conversely, the free throw rate has generally decreased, reflecting a shift away from reliance on free throws. This could be a consequence of maintaining a faster pace of play or focusing on scoring from behind the arc. Figure 1a, 1b, and 1c visualize the trends in these metrics.

Looking at broader metrics such as offensive and defensive rating reveals interesting insights. Generally, championship teams ratings on both sides of the court were better than average, but there were some exceptions. Blue dots lower than red dots in Figure 2a and blue dots higher than red dots in Figure 2b exemplify these exceptions. However, a championship team with a peculiarly weak rating on one side of the court would compensate for that rating with an exceptional rating on the other side of the court, therefore achieving a high overall net rating (difference between offensive and defensive rating). After graphing net rating, it became clear that this metric was the most reflective of a championship team's abilities. Figure 2c displays consistent separation of blue and red dots in Figure 2c.

Besides net rating, what's important to note is that efficiency is an extremely important metric for success in the league. Championship teams not only attempt more three-pointers (Figure 3a), but convert them at increasingly higher rates compared to non-championship teams (Figure 3b). Effective field goal percentage (eFG%) also distinguishes championship teams, who generally hold higher percentages compared to the league average, with few exceptions shown in Figure 3c. This metric, which weights field goal efficiency by the amount of points each shot is worth, reflects a team's ability to maximize scoring opportunities through high-percentage shots, whether from inside the paint or beyond the arc.

The age and basketball experience of players, which generally go hand in hand, are significant factors in championship success as evidenced by Figure 4a. While the average age of all teams has decreased over time due to increasing development of young players, championship teams consistently feature a higher average age than their non-championship counterparts. This suggests that veteran experience and leadership are crucial elements for achieving success in the postseason.

Dialing in on more specific aspects of the game, team assists are found to be generally higher among championship teams, particularly in recent years, which highlights the importance of ball movement in creating scoring opportunities and breaking down opposing defenses. Championship teams also maintain a lower opponent free throw rate, indicating disciplined defensive execution that limits opponents' scoring. Figures 4b and 4c visualize these trends.

Metrics in blocks, rebounds, and turnovers present less consistent pictures. While some championship teams achieved significantly higher blocks per game, it was not a definitive factor for success. It is likely that defensive cohesion, the ability to contest shots effectively without necessarily blocking them, or other factors outside of a team's defense may be more critical in many cases. Offensive and defensive rebounding percentages tell a similar story, with championship teams sometimes outperforming non-championship teams in these areas. The varying turnover rates suggest that teams can succeed with different levels of ball security. While minimizing turnovers is generally beneficial, it is not the sole determinant of championship success.

Finally, there were several factors that we concluded did not hold much influence in defining a team's championship success. The first is supremacy in win percentage; having the absolute highest regular-season win percentage does not necessarily predict a championship victory. The top win percentage team has only won the championship ten times in the past three decades, highlighting the unpredictable nature of the playoffs and the additional dynamics that come into play during the postseason, such as matchup-specific strategies, environment, and player health. For example, many teams that were very successful in the regular season have fallen short in the playoffs due to untimely injuries for key players, which is not accounted for in team metrics.

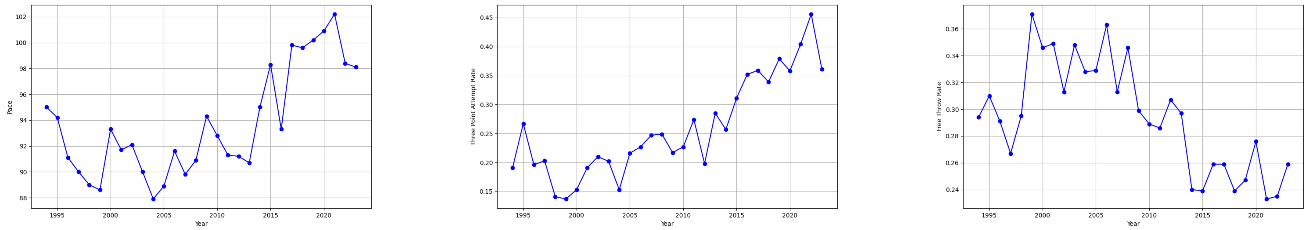
Improvement in a team's performance from one season to the next does not strongly correlate with winning a championship. In the past 30 years, only four times has the championship team been a "top five" improver from the previous season, indicating that more consistent improvement and excellence are more critical than short-term improvement. This finding highlights the importance of building a solid foundation and maintaining a high level of performance over multiple seasons. The most recent champion in 2023, the Denver Nuggets, consists of a roster carefully built across many seasons. This team saw multiple playoff stints in recent years, up until their most successful championship run last season.

Strength of schedule and attendance figures are unconvincing in predicting championship success over the years. Some championship teams have faced more challenging schedules, while others have benefited from easier ones. However the ultimate determinant of winning a championship is performance in the playoffs, where teams face off against the best in the league. And while championship teams generally enjoy higher game attendance, this could be influenced by factors such as market size, overall team success, and the excitement generated by a winning team. It is very possible that successful teams draw more fans, creating a positive feedback loop of support and performance. Higher attendance itself does not necessarily drive championship outcomes.

Overall, our findings in exploratory data analysis, supported by various visualizations, provide an all-around understanding of the attributes that contribute to NBA championship success. While certain trends and metrics are indicative of potential success, the path to a championship is influenced by a group of factors that interact in complex ways throughout the season and into the playoffs.

IV. Proposed Methodology

Our methodology begins with picking a model that is capable of comprehending the complexity of our datasets, while remaining adaptive to unseen data. We considered three machine learning algorithms: multiple linear regression, Bayesian probability networks, and deep neural networks (DNNs). Each of the mentioned algorithms possess their own unique strength and limitations.

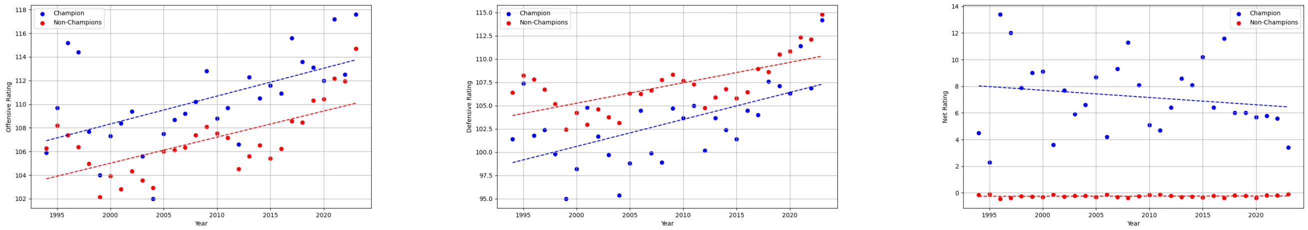


(a)

(b)

(c)

Fig. 1: Line chart for (a) pace; (b) three point attempt rate; and (c) free throw rate

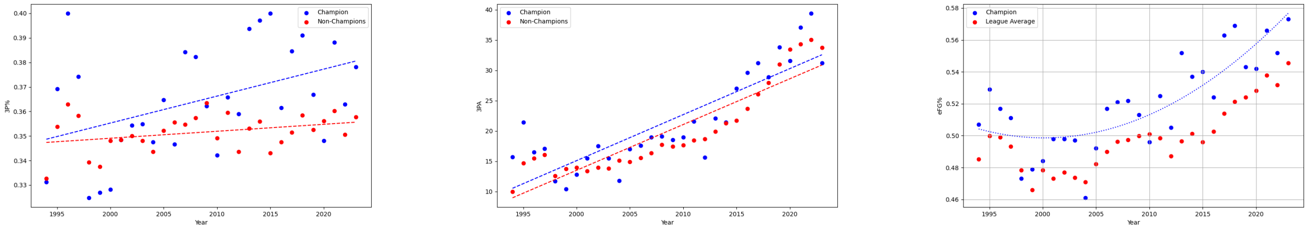


(a)

(b)

(c)

Fig. 2: Scatter plot for (a) offensive rating; (b) defensive rating; and (c) net rating

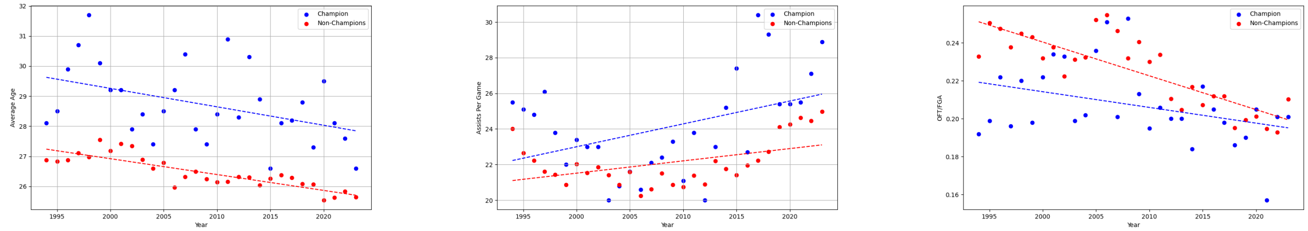


(a)

(b)

(c)

Fig. 3: Scatter plot for (a) three point percentage; (b) three point attempt rate; and (c) effective field goal percentage



(a)

(b)

(c)

Fig. 4: Scatter plot for (a) average player age; (b) assists per game; and (c) opponent free throws per field goal attempted

Multiple linear regression, while a straightforward statistical learning model, is not well-suited for capturing complex relationships in the data necessary for sports analytics. Multiple linear regression represents the linear relationship between a set of independent predictor variables (the input features) and a dependent outcome variable. The advantage to linear regression is that it is highly interpretable, which helps in understanding how individual variables influence the prediction. It is also relatively computationally inexpensive. However, the drawback to linear regression models is the assumption of linearity, which is often not the case in the realm of sports as we will observe later in the model evaluation.

Bayesian probability networks represent the championship likelihood as a combination of conditional prior probabilities through a direct acrylic graph (DAG). Probability networks are excellent in modeling complex interaction between variables and integrating prior knowledge, which is essential for sports prediction models in which future outcomes depend on past trends. However, the disadvantage to Bayesian models is that they require a thorough understanding of the network and underlying prior distributions, which can be challenging to obtain due to the model's randomness. Additionally, they are computationally expensive.

Deep neural networks (DNNs), on the contrary, possess the ability to handle the complexity and randomness of sports. DNNs excel at identifying patterns that may not be evidently apparent and are capable of learning features at a higher abstraction level, which is paramount for capturing the subtle nuances of the game of basketball. The capability of DNNs to self-learn these nuances with minimal feature engineering and tuning place it as the optimal choice for our model. In spite of its computational complexity, advancements in GPU technology specifically designed for deep learning models such as DNNs provide an ideal platform to train highly complex data in fast iterations. The limitation of neural networks, however, is their lack of interpretability compared to simple models. However, since we are dealing with complex datasets, this is not a requirement that we need to consider when choosing our learning algorithms.

Following the exploratory data analysis (EDA) to identify the most important features, the training process begins with a thorough preparation and preprocessing of data collected from 1994 to 2023. This data is synthesized into a single DataFrame, creating a comprehensive source for in-depth analysis. During preprocessing, significant features from the EDA are selected, while insignificant predictors and categorical data are excluded. We employed data normalization using the standard scaler to standardize all continuous variables. Normalization is essential for optimal performance in neural networks to ensure uniform input scales across all features, removing bias, maintaining constant variance, and most importantly, enhancing the model's accuracy.

The architecture of our DNN model begins with an input layer consisting of all the selected input features. The input layer is followed by 3 hidden layers, each containing 256 units with a ReLU activation function. The ReLU functions allow the model to explore the non-linearity and non-apparent correlations of the dataset. The output layer consists of a single unit signifying the likelihood of that particular team winning the NBA championship. This output node layer follows a sigmoid activation so that the predicted outcome is between 0 and 1, similar to our target championship values. The network is then compiled using the Adam optimizer, coupled with a mean squared error loss function, to streamline the training process by effectively minimizing errors and ensuring convergence. A diagram of our neural network can be found in Figure 5.

To prevent overfitting, a 10-fold cross-validation strategy is implemented additionally. This method enhances the model's generalizability, as it tests the network's performance across different data subsets, ensuring its robustness and adaptability to unseen data, while providing a reliable estimate of the model's overall performance. Each fold involves a cycle of training on selected data subsets and validating on others, with continual monitoring and adjustment of error metrics to fine-tune the model iteratively. We run the model at each iteration for 100 epochs with a batch size of 64, allowing for gradual learning and efficient backpropagation. This approach helps in progressively refining the model's weights to better identify and adapt to the underlying patterns in the data without overfitting. At each iteration, the model and the mean squared error (MSE) between the predicted and actual values of the test data set are stored to compare the errors of the models at each fold. Finally, the model with the lowest MSE value is chosen as the best model for the championship likelihood predictions.

Conclusively, the best-performing model is utilized to forecast the NBA Champion. For unseen data, we follow the same data pre-processing steps of normalization and feature selection. We then feed in the new data to the neural network model where it will calculate the likelihood of each particular team winning the NBA championship for the respective season. Thus, the team with the highest likelihood in that season is predicted to be the NBA champion for that particular year. These probabilities provide valuable insights in depicting potential outcomes. To further evaluate the model, we test the selected best model on historical data from 1994 to 2023, and track how many of the historical NBA champions our model is able to correctly predict in order to get an accuracy score. Thus, our comprehensive methodology for predicting the NBA champion is not only underscored by our model's predictive accuracy, but also its applicability to new unseen datasets.

V. Experimental Results and Evaluation

Following the selection of the best model from our cross-validation, we tested the model on historical data spanning from 1994 to 2023. In this evaluation, our model achieved a remarkable 90% accuracy, predicting 27 out of the 30 historical champions correctly, failing to predict the 2005 San Antonio Spurs, the 2016 Cleveland Cavaliers, and 2022 Golden State Warriors, instead predicting the Phoenix Suns, Golden State Warriors, and Boston Celtics, respectively.

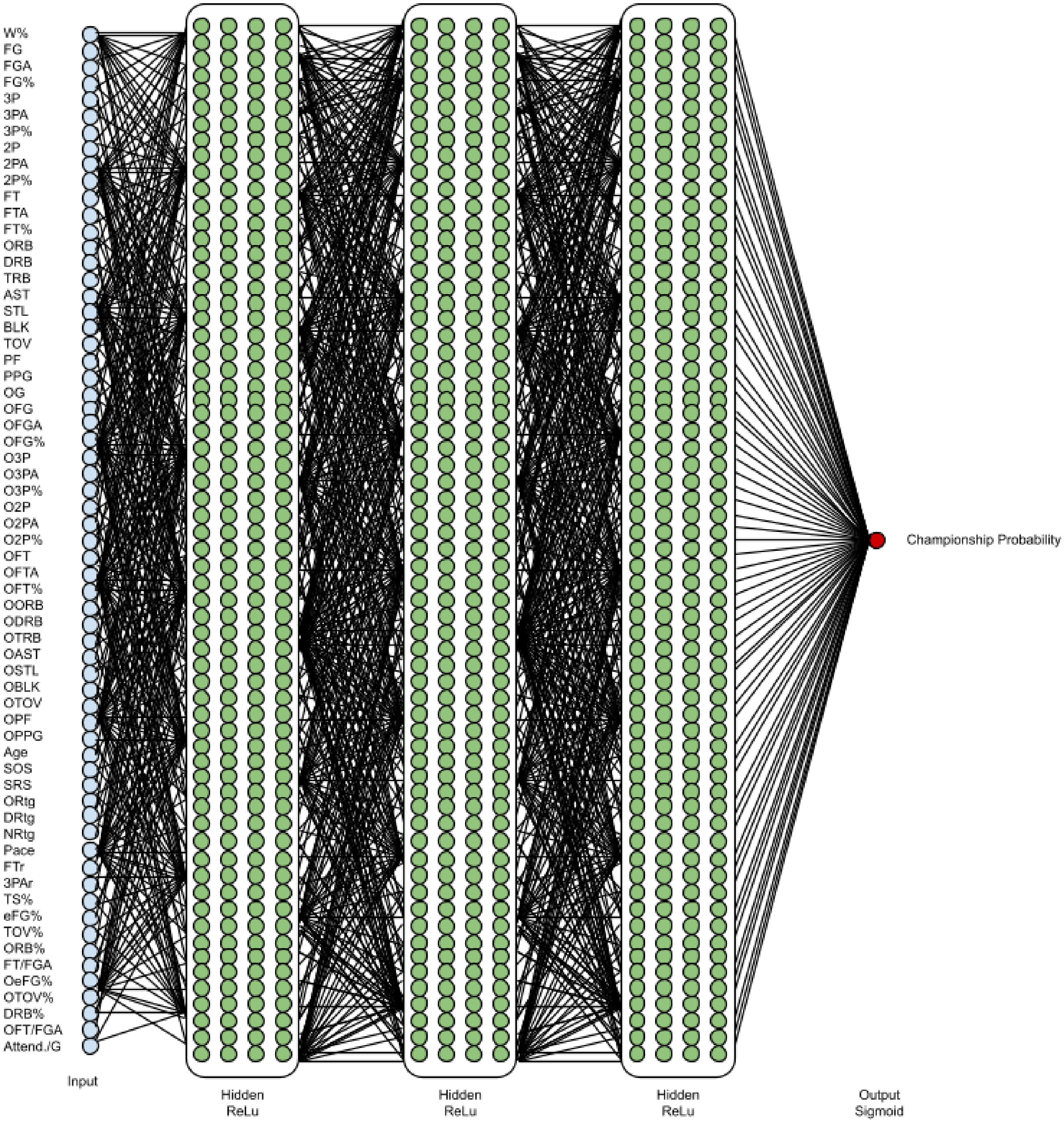


Fig. 5: Deep neural network to predict championship probabilities

While a 90% accuracy is impressive on its own, the errors in the historical prediction do not undermine our model since each of the incorrectly predicted teams were statistically and media favorites to win the championship. In particular, the 2016 Golden State Warriors, despite securing the most regular season wins in a season and being clear cut favorites for the NBA title, fell short to the Cleveland Cavaliers in 7 games in the NBA finals. These errors are more of a testament to the stochastic nature of the NBA rather than the capabilities of our model.

Although the final evaluated model predicted the Boston Celtics with 90% confidence as the 2024 NBA Champions, the model regularly predicted between the Boston Celtics and Minnesota Timberwolves during the training iterations. This indicates that there is variability in the model, the model remains consistent in the likelihood of specific teams to win over others. Additionally, it illustrates the stochastic nature of neural network architectures. Despite hyperparameters remaining constant, the model would display variability between the two prediction outcomes, underscoring the complexity of the back propagation techniques used to achieve varying local minima in the gradient

In spite of the stochasticity of the model training, optimal training hyperparameters were crucial to achieving the best prediction model. During the multiple training iterations that we performed, we observed the performance of models with different hyperparameters. Namely, we tested between 50 and 100 epochs, a batch size between 32 and 64, and hidden layer sizes of 128, 256, and 512 units. After extensive testing between various combinations of these hyperparameters, we concluded that the optimal model configuration consisted of 100 epochs, a batch size of 64, and hidden layer size of 256 units. This setup struck the best balance between computational efficiency and predictive accuracy, confirming that appropriate tuning of hyperparameters is crucial for maximizing model performance.

In the exploratory data analysis (EDA) stage, we identified a list of 12-20 significant features, which we later used to select the features for the deep neural network model. However, when training the model with only the features that we deemed significant from the EDA stage, our model obtained an average accuracy of around 67%. However, upon expanding the scope of the input features to include all of the numerical features, the accuracy of the model improved significantly, and we were able to consistently train models that obtained an accuracy of over 80%. This improvement suggests that while certain features may not display any significance in direct correlation to the championship outcome, it may still hold some indirect or non-linear correlation to the outcome variable. This complexity in the relationship of the input features to the outcome serves as a testament to traditional statistical techniques not sufficing for accurate predictions, and it further validates our choice to use a deep neural network model over a regression model.

Despite the high predictive accuracy and consistent likelihood values based on the statistical learning model, the ongoing 2024 NBA playoffs highlighted the unpredictability of the NBA landscape. Based on the model likelihood values, the Minnesota Timberwolves were expected to win the Western Conference. However, the Dallas Mavericks took an unexpected leap during the postseason despite being given a 2% likelihood to win the championship, and defeated the Minnesota Timberwolves to advance to the NBA finals. Nonetheless, such occurrences are not unprecedented in the game of basketball. In last year's postseason, the Miami Heat, to everyone's surprise, took down many Eastern Conference title favorites to advance to the NBA finals. Similarly, in 2016, despite having arguably the greatest regular season in NBA history, the Golden State Warriors fell short in the NBA finals to the Cleveland Cavaliers as a result of an all-time great performance by LeBron James and Kyrie Irving. Such outcomes exemplify the fact that even the most sophisticated models cannot account for the unforeseen variables that may rise. At the time of writing, the Boston Celtics lead the 2024 NBA Finals 1-0 against the Dallas Mavericks, 3 games away from securing the championship. Using probabilistic calculations where each team is equally likely to win any given team, this yields a 67% chance of winning the 2024 NBA Championship in favor of the Boston Celtics. This serves a real-world benchmark for our model's predictive accuracy. Using only data from the regular season, our model was able to accurately predict the Boston Celtics, who still remain as heavy favorites for the 2024 NBA champion, further validating our model's capabilities and credibility in the sports realm.

In conclusion, the detailed evaluation of our neural network model illustrates its potential as a dynamic and powerful analytical tool in sports. It highlights the necessity for continuous refinement and adaptation based on new data and evolving team dynamics to maintain its relevance and accuracy. Nonetheless, its capabilities as an accurate prediction model are apparent, being able to adapt to unseen data, and make accurate predictions. Moving forward, the insights gained from this project will inform further improvements and deepen our understanding of the key factors that drive outcomes in competitive sports, paving the path for further refinement to improve the model's credibility as a powerful analytical sports tool.

VI. Conclusion and Discussion

The purpose of this project is to predict the NBA championship for the upcoming season using machine learning techniques. Using statistics and metrics, we wanted to look deeper into the NBA and find what trends and metrics truly lead to the winning of the NBA championship and also if we can predict it for the future season. Our comprehensive approach involved sourcing and preprocessing datasets from Basketball-Reference, conducting exploratory data analysis (EDA) to identify key trends, and developing a predictive model based on the insights gained.

Some of the key findings from our project were from the twenty data analysis questions we as a group created for our exploratory data analysis (EDA). From our EDA, we found multiple trends of NBA teams and how certain statistics stand out when looking at NBA championship teams and NBA non-championship teams. Some statistics that stood out were win percentage, field goal percentage, field goals attempted, 3-point percentage, free throw percentage, offensive rebounds, defensive rebounds, assists, steals, blocks, points per game, opponent points per game, age, offensive rating, defensive rating, net rating, free throw rate, true shooting percentage, effective field goal percentage, and pace. Using these metrics, we were able to answer our question of "which statistics can be directly correlated with winning the NBA championship.

The main finding from the project, who would win the 2023/2024 NBA Championship, the Deep Neural Network model predicted the Boston Celtics to win the NBA Championship this season with a 37.86% likelihood. The model had a 90% accuracy of predicting the past NBA Championships from the year 1994 all the way to 2023. This was the model that had the highest accuracy score for predicting the past championships from all our training iterations, as well as the one that made the most sense to us as a group when looking at the individual team probabilities of winning the NBA Championship.

Our NBA Championship predictor contributes to the field of sports and practical applications by enhancing possible decision making scenarios for teams and coaches. One of the most important contributions of projects like these are the fans' benefits. This is due to increased engagement and entertainment through access to predictive analytics. These predictive analytics can also spark more discourse to media broadcasters for even more engaging content. Another big market that these predictive models can reach is the sports betting market. In this market, predictive models offer data-driven insights for better betting strategies and market efficiency. Finally, the teams themselves can benefit from the creation of further advanced predictive models. Team executives can make informed decisions about player transactions and salary cap management and build rosters destined for success. Overall, NBA Championship predictors drive innovation in sports analytics, deepening the understanding of the game and enhancing various aspects of the sports industry.

One of the biggest limitations from the project was the individualism factor in the NBA playoffs. Certain players have the "X" factor when it comes to the biggest games of the season, and these individual statistics could not be accounted for in our predictive models. Including these "x-factor" players and how they perform on the biggest stage in the world could have driven us to change our model or even choose a completely different model. Due to us not including individual players statistics, our model could have some skewed probabilities for certain teams. A prime example for this would be the Dallas Mavericks. They have two of the best examples of such players in Luka Doncic and Kyrie Irving. These two players can take control of games on their own and fuel their team to important wins in the playoffs, and they demonstrated this against the Los Angeles Clippers, Oklahoma City Thunder, and the Minnesota Timberwolves, as they were the underdogs in each playoff series. If we used these facts in our predictive model, the Dallas Mavericks would most likely have a higher probability of winning the NBA Championship than 2.26%, which is the likelihood that our model gave them.

For the future, our model could be significantly enhanced by incorporating a few key improvements. Firstly, by taking into account "x-factor" players—those individuals whose exceptional performance and unique skills can dramatically influence game outcomes—we could refine our predictions to account for these critical contributors. Including metrics that capture individual player impact, such as clutch performance and playoff experience, would provide a more accurate prediction model. Secondly, integrating real-time data updates would allow our model to dynamically adjust to current season developments, such as injury history, trades, and team form fluctuations, ensuring that our predictions remain relevant and accurate throughout the season. Lastly, expanding our predictive capabilities to forecast championships for multiple future seasons would involve developing longitudinal models that consider trends and player development trajectories over time. This could include analyzing player career arcs, team rebuilding phases, and historical performance trends. By incorporating these enhancements, our project would evolve into a more sophisticated tool, offering deeper insights and more accurate predictions, ultimately providing greater value to teams, coaches, analysts, and fans alike. These advancements would not only improve the immediate predictive power of our model but also establish a trustworthy framework for long-term strategic planning in the NBA. However, the drawback to these advancements are the increased computational power necessary to account for all these factors.

Sports networks, analysts, broadcasters, managers, and fans all predicted the NBA Championship this year to be the Boston Celtics, which aligns with our model predictor. This gave our group a positive outlook about how our model could be a success in predicting the NBA Championship. Our model works differently from models that can be found online and around the internet. Some projects online used different models such as Logistic Regression (LR), Linear Discriminate Analysis (LDA), Support VectorMachine (SVM), K-Nearest Neighbors (KNN) and Classification and RegressionTree (CART), while we decided to use Deep Neural Networks (DNN). Deep Neural Networks are distinct from these other machine learning models and better for our project due to the complex architecture and ability to model the non-linearity of the NBA. Our deep neural network was able to capture the intricate patterns through multiple layers of connected neurons. This allowed our project to handle high-dimensional data and complex, non-linear decision boundaries more effectively than the other models. One of the similarities of our project to other NBA Championship predictors was the idea that certain statistics can lead to teams having a higher probability of winning the NBA Championship.

The model we created has a multitude of strengths and weaknesses. One of the biggest strengths of the project is the high accuracy. For our model to correctly predict the NBA Championship 90% of the time from 1994 all the way to 2023 is testament to how accurate the model is. As we also had twenty features to include in our model, the handling of high dimensional data was another strength that our model had. As the NBA is a very unpredictable league and environment, the handling of non-linear relationships was also another strength that our model dealt with very well. A weakness of our model is the individual probability scores that teams could get in winning the NBA Championship. Although the main goal is to predict the final winner of the NBA, the individual scores can give our model a negative outlook. An example would be the probability scores for teams to win this year's NBA Championship. The Oklahoma City Thunder got a probability rating of 0.16% despite holding the best record in the Western Conference. This score is just 0.01% more than the Orlando Magic. Different sports analysts and fans alike could tell just by looking at the team rosters and statistics, that the Oklahoma City Thunder should have two to three times the probability score of the Orlando Magic because of the difference in quality of the teams and the statistical inequalities between the Thunder and the Magic. Although the Oklahoma City Thunder aren't the clear favorites to win the championship, they did have a deep run and met their high expectations. This has to do with the unpredictability of the NBA. These nuances could make our model to be seen as "unreliable" or "inconsistent". Another weakness of our model is its adaptability. Although this model's goal is to predict the championship for this next season, to use it in the future, we would need to add in all of the season's statistics to the raw data file, and then rework the model to include that season's statistics. This can be computationally expensive and also take a lot of time.

All in all, we believed our project to be a success. From the beginning stages of the exploratory data analysis and feature selection to the model testing and deployment, this project has resulted in a fulfilling NBA Championship prediction. During the EDA stage, we cleaned and analyzed the historical data of each and every NBA team from 1994 to 2024 and identified the key statistics and metrics that would lead teams to winning the NBA Championship. Feature selection allowed us to pinpoint the most predictive variables, refining our model for greater accuracy. Rigorous testing ensured that our predictions from the model were reliable and the final deployment provided us with an intuitive interface for other users to engage with the findings of our model. Thus, we conclude that this project not only highlights the power of machine learning and deep neural networks in predicting the NBA Championship outcomes but also serves as indication of our team's dedication, analytical skills, and innovative approach in the realm of machine learning.

VII. Github Link and Project Roadmap

Github: https://github.com/amubaraksyed/NBAChampion
Demo: https://youtu.be/A8Epfwda0qk

Project Roadmap:
April 22 - Submit 1-Page Project Topic Selection
April 28 - Data Extraction and Cleaning
May 3 - Exploratory Data Analysis
May 5 - Mid-Quarter Progress Report
May 19 - Model Training, Evaluation, Results
May 24 - Complete Frontend and Backend Interface
May 31 - Model Deployment and Testing
June 2 - Project Report Rough Draft
June 6 - Project Report Final Draft
June 8 - Submit Project Report, Source Code, Demo

All team members contributed to the brainstorming of project topics and goals, which was reflected in the one-page report. The mid-quarter progress report was developed between all group members. The exploratory data analysis, involving brainstorming questions and answering them through Python graphs, was equally divided among the team members.

Aadhil was primarily responsible for data extraction and initial cleaning, as well as model training and evaluation. Jaynor handled the development of the front end to deploy our machine learning model. Dylan and Zareb contributed significantly to the project report, including the literature review and a major cleanup of the exploratory data analysis conducted by individual members.

Throughout the project, no aspect was completed by a single member without the corroboration, modification, or review of the other team members.

References

[1] Bunker, R., Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1), 27-33.

[2] Digital, B., Bryant, R., Repository, D., Houde, M. (n.d.). Predicting the Outcome of NBA Games Predicting the Outcome of NBA Games Part of the Databases and Information Systems Commons, and the Data Science Commons. https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1000&context=honors_data_science

[3] Haghighat, M., Rastegari, H., Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, 2(5), 7-12.

[4] JakeAllenData. (2024, March 4). Predicting The NBA Champion With Machine Learning. Medium. https://allenjake440.medium.com/predicting-the-nba-champion-with-machine-learning-25e3a45a82f9

[5] Liu, H., Xu, D., Li, Y. (2018). Predicting NBA playoff outcomes using machine learning techniques. *Journal of Sports Analytics*, 4(1), 65-77.

[6] Liu, Sean. (2021). Predicting NBA Playoffs Using Machine Learning. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/349646430_Predicting_NBA_Playoffs_Using_Machine_Learning

[7] Schumaker, R. P., Solieman, O. K., Chen, H. (2010). *Sports data mining. Springer Science Business Media.*

[8] TheJK. (2022, August 25). I will predict the 2023 NBA Champion using Machine Learning. Medium. https://thejk.medium.com/i-will-predict-the-2023-nba-champion-using-machine-learning-5e8df072059d

[9] Yan, S., Lucey, P., Morgan, S., et al. (2020). Predicting the outcomes of NBA games using deep learning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 1459-1468.

[10] Yang, J., Lu, C.-H. (n.d.). Proceedings of Artificial Intelligence and Machine Learning for Engineering Design AI ML for ED Spring, 2012, PREDICTING NBA CHAMPIONSHIP BY LEARNING FROM HISTORY DATA. https://static1.squarespace.com/static/51bb9790e4b0510af19ea9c4/t/51bf8760e4b0356bbe2911f1/1371506528092/nba_champ_predict.pdf