UNIVERSITY OF STRATHCLYDE

DEPARTMENT OF
MATHEMATICS AND STATISTICS

Beyond the Arc: Predicting NBA Games Through Machine Learning

by

Euan Smith

202002654

BSc Hons Mathematics, Statistics and Economics
2023/24

Statement of work in project

The work contained in this project is that of the author and where material from other sources has been incorporated full acknowledgement is made.

Signed ……

Print Name ……

Date ……

Supervised by Louise Kelly

Contents

1 Introduction

Predicting the outcomes of NBA games and the consequent standings of teams is a challenge that statisticians and basketball enthusiasts have grappled with for years. The allure of the task lies in its complexity and the dynamic nature of the sport and the National Basketball Association (NBA) provides a rich bank of data for prediction.

The NBA has been an extensive collector of game-related data, making it an ideal candidate for statistical analysis. The vast amount of data available allow for a nuanced exploration of the factors that contribute to a team's victory, thereby facilitating the creation of a predictive models.

The implementation of the three-point line in the 1979-80 season marked a significant turning point in basketball strategy. Initially regarded as a risky and unconventional play, the three-point shot has evolved to become a central element of basketball offense. This evolution has been gradual, reflecting the sport's adaptability and the innovative mindset of its participants. The analytics movement in basketball, spearheaded by individuals like Daryl Morey, placed a spotlight on the efficiency of the three-point shot and led to a league-wide re-evaluation of offensive strategies. The Houston Rockets, under Morey's leadership, exemplified this shift as they often led the league in three-point attempts. Today, the NBA is characterised by the prevalence of the three-point shot, with players such as Stephen Curry and Klay Thompson, the famed "Splash Brothers" of the Golden State Warriors, revolutionising the guard position along with winning 4 NBA championships together (Freitas 2021).

The Elo rating system is measure of a team's competitive calibre. Teams start with an initial Elo rating which then evolves through their performance in games, with the rating's sensitivity to match outcomes dictated by a carefully chosen K-factor, for the NBA balancing the need for ratings to respond to recent performances without excessive volatility. This system has the added advantage of incorporating the margin of victory, providing a more nuanced understanding of each game's context.

In this report, we will be looking at a large portion of NBA history, with team statistics from the 1983/84 season all the way up until the 2020/21 season, with the objective of employing logistic regression to predict the outcomes of the 2020/21 NBA season. Logistic regression has been selected for this study due to its effectiveness in analysing binary outcomes,

specifically, the wins or losses of NBA games. These outcomes are influenced by a variety of predictor variables, reflecting the diverse and complex nature of team performance, including Elo rating. The model aims to break down the patterns in team performances and predict game outcomes with a high degree of accuracy.

2 Overview of NBA

The National Basketball Association (NBA) is a professional basketball league in North America, featuring 30 teams (29 in the United States and 1 in Canada). It is renowned for hosting some of the world's most talented and highly paid athletes and generally considered the best, highest-level basketball league in the world.

The NBA is organised into two main conferences: the Eastern Conference and the Western Conference, each subdivided into three divisions. The regular season, spanning from October to April, entails each team playing 82 games. Within this there is an All-Star Weekend which serves as an interlude for the regular season and is a celebration of the league's top talent in various exhibitions and competitions. This setup ensures a mix of divisional, conference, and inter-conference matchups, leading to a comprehensive and competitive season.

Some things to know about the NBA and basketball, in general, are:

- NBA games are divided into four 12-minute quarters, with overtime periods of 5 minutes if the game is tied at the end of the fourth quarter.
- Each team has five players on the court. Substitutions can be made during stoppages in play.
- Points are earned through field goals (two points, or three points for shots beyond the three-point line) and free throws (one point each).
- An assist is credited to a player who passes the ball to a teammate in a way that leads to them scoring a field goal.
- A free throw is a shot awarded after certain fouls, taken from the free-throw line without opposition.
- A rebound is when a player retrieves the ball after a missed shot attempt. There are two types of rebounds: offensive and defensive. An offensive rebound is when the

attacking team regains possession, whereas a defensive rebound is when the defending team gains possession following the opponent's missed shot.

- A block is a defensive play where a player legally deflects a shot attempt, preventing it from going into the basket.
- A steal is when a defensive player legally takes the ball away from an opponent, gaining possession.

Fouls are called for illegal physical contact, with players being limited to six personal fouls per game before disqualification. Violations, including traveling, double dribbling, and shot clock infractions, lead to a change in possession or loss of scoring opportunity.

The postseason begins with the Play-In Tournament, determining the final two playoff spots in each conference for teams placed 7th to 10th. The playoffs then proceed in a best-of-seven format across the First Round, Conference Semi-finals, Conference Finals, and then the NBA Finals, where the champions of each conference fight for the bragging rights to the NBA championship, the most prestigious championship in all of basketball, all this information so far can be found on the NBA's official website from the official NBA Guide which is annually published by the league and provides detailed information on the league's format, rules, and organisational structure.

3 Methods

3.1 Elo Rating System

The Elo rating system, created by physicist and chess player Arpad Elo, has become a standard method for calculating the relative skill levels of players in various types of competitive games. Elo's central premise was to construct a statistical representation of a player's strength that could be updated in response to game outcomes. The system calculates the expected probability of winning based on the rating difference between the competitors, using a logistic probability distribution (Elo 1978). In chess, the Elo rating system offers a solution to rating players skill level. The system's flexibility allows it to reflect the dynamic nature of a player's skill, rising with wins and falling with losses. This simplicity and adaptability have led to the system's wide acceptance and application beyond chess to other areas, including various sports (Marveldoss 2018; Hvattum & Arntzen, 2010).

The mathematics behind the Elo rating involves calculating the expected outcome for each team or player before a game starts, based on their current ratings. This expected score is then compared to the actual result of the game to adjust the ratings. The actual update to a team's or player's rating is calculated using the formula:

$$R_{i+1} = R_i + K(S - E)$$

Equation 1.1

In equation 1.1; R_{i+1} is the new rating, R_i is the current rating, K is the K-factor, which determines the sensitivity of the rating system to a single game result, S is the actual outcome of the game (1 for a win and 0 for a loss), E is the expected outcome of the game, calculated from the ratings of the participants before the game.

The expected outcome is calculated using the from the formula:

$$E = \frac{1}{1 + 10^{\frac{R_O - R}{400}}}$$

Equation 1.2

In equation 1.2; R_O is the current rating of the opponent, E is the expected outcome of the player/team of interest. FiveThirtyEight is a sports blogging website which has created their own application of the Elo rating system to the NBA which preserves the foundational principles of Elo while incorporating NBA-specific modifications, details of this can be found on their website in the article 'How We Calculate NBA Elo Ratings'. The core concept of the rating system remains the same, with teams ranking increasing with wins and falling with losses. These ratings cover regular-season, play-in and playoff games, drawing from the comprehensive historical data source "Basketball-Reference.com".

FiveThirtyEight's system, like the chess system, is inherently zero-sum where points gained by one team after a game are equivalently lost by the opponent. FiveThirtyEight's model introduces a K-factor, optimised for the NBA at a value of 20. This optimised K-factor addresses the NBA's high frequency of games but still allows for a relatively swift response to shifts in team quality. It suggests a higher weighting for recent performance. Such calibration enables the model to promptly recognise genuine changes in team quality without excessive volatility that might result from an overreactive K-factor.

FiveThirtyEight also factors in the importance of home-court advantage, set at a constant 100 Elo points. This constant reflects empirical evidence, although the actual advantage has varied throughout NBA history (Nevill & Holder 1999).

Furthermore, a margin of victory, multiplier is integrated into the system, with a mechanism to account for diminishing returns on larger margins, reflecting the principle that the difference between a 5-point win and a 10-point win is more significant than between a 25-point and a 30-point win. This sophistication in the model allows it to discern between hard-fought close games and more definitive victories, providing a more detailed perspective on team performance (Kovalchik 2020):

$$G = \frac{((MOV + 3)^{0.8})}{(7.5 + 0.006(R_W - R_L \pm Home\ Advantage))}$$

Equation 1.3

In equation 1.3, MOV is the margin of victory of the winning team, R_W is the current rating of winning team, R_L is the current rating of losing team, Home Advantage (100) is added or subtracted depending on location of winning team.

This multiplier is integrated into equation 1.1 as such:

$$R_{i+1} = R_i + KG(S - E)$$

Equation 1.4

Finally, year-to-year carry-over is an integral part of FiveThirtyEight's adaptation. Unlike resetting ratings at the beginning of each season, a portion of the previous season's ratings is carried forward. This fraction is higher for the NBA than for NFL ratings in FiveThirtyEight's models, underscoring the greater consistency in team performance from year to year in basketball. For instance, a team's Elo rating at the start of a new season is a weighted combination of its final rating from the previous season and the league average:

$$R_{s=i+1} = (0.75)R_{s=i} + (0.25)1505$$

Equation 1.5

The inclusion of factors specific to the NBA, such as accounting for expansion teams by starting them with a lower base rating of 1300 and adjusting the league average Elo slightly higher than 1500 to balance the ratings, demonstrates the careful calibration required for the Elo system to accurately reflect the nuances of the league. These parameters ensure that

new teams are appropriately rated while maintaining the integrity and comparability of the Elo ratings across different eras.

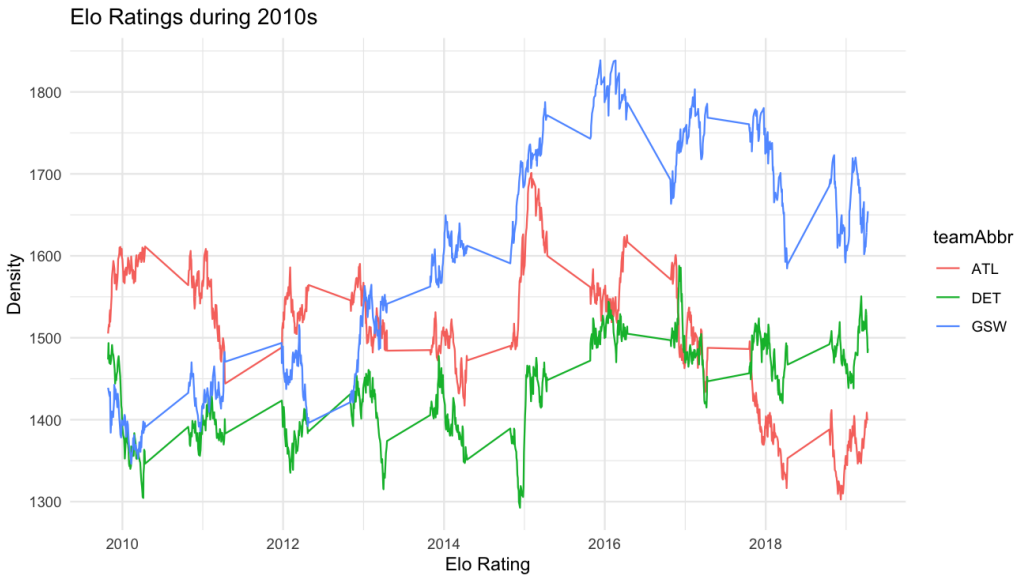

Figure 1 - Elo Ratings of NBA teams throughout 2010s

Figure 1 clearly depicts that in 2016 during the Golden State Warrior's record-breaking season of 73 wins and 9 losses they had an extremely high Elo rating. Therefore, this gives evidence of how Elo is a good indicator of how strong a team is.

3.2 Logistic Regression

Logistic regression is a fundamental statistical method for binary classification which traces its roots to the logistic function and serves as a bridge between linear predictors and binary outcomes. At the heart of logistic regression lies the logistic function:

$$\sigma(z) = \frac{1}{1+e^{-z}},$$

Equation 2.1

which maps any real-valued number into the (0, 1) interval, making it interpretable as the probability of the dependent variable belonging to a particular category. This characteristic is crucial for binary outcome prediction, such as pass/fail or yes/no, scenarios. The relationship between the log odds of the outcome and the predictors is linear:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K,$$

Equation 2.2

p is the probability of the outcome of interest, β_0 is the intercept and β_1, …, β_K are the coefficients for the predictors X_1, …, X_K.

The estimation of these coefficients is typically achieved through maximum likelihood estimation (MLE), a method that iteratively tests various values for β to find the set that maximises the likelihood of observing the sample data (Cox, 1958).

In essence, logistic regression's mathematical underpinnings ensure that it remains a powerful tool for binary classification across various domains, from medicine, political science and sports analysis where understanding the probabilistic relationship between predictors and categorical outcomes is vital (Hosmer Jr, Lemeshow & Sturdivant, 2013).

3.3 Model Evaluation Techniques

3.3.1 Confusion Matrix

The confusion matrix is a tool commonly used in classification tasks to visualise the performance of a predictive model. It displays the frequency of each class based on the true class versus the predicted class. Each entry in the matrix denotes the number of predictions made by the classifier compared to the actual outcomes in the data, broken down into true positives, true negatives, false positives, and false negatives (Pearson, 1904).

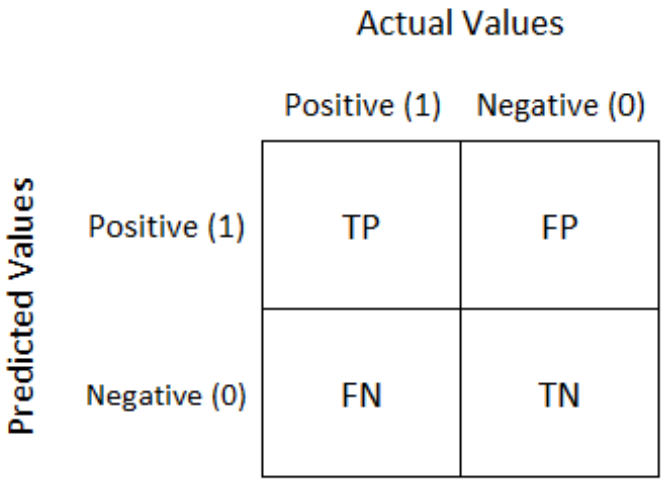

Figure 2 - Confusion Matrix Visualisation (Image courtesy: My Photoshopped Collection)

Sensitivity measures the proportion of actual positive cases that are correctly identified by the classifier. It is an indicator of a model's ability to detect positive instances effectively and

it is particularly important in areas where failing to identify positive cases can have serious consequences, such as when diagnosing disease. It is calculated as such:

$$Sensitivity = \frac{TP}{TP + FP}$$

Equation 3.1

Specificity assesses the proportion of actual negatives that are correctly identified and is crucial for situations where it is essential to be certain of a negative prediction. For instance, in the context of spam detection, a highly specific model ensures that non-spam emails are not incorrectly classified as spam. It is calculated as such:

$$Specificity = \frac{TN}{TN + FN}$$

Equation 3.2

Accuracy is the most comprehensive performance measure, and it is simply a ratio of correctly predicted observation to the total observations. It gives an overall insight of the success rate of a model but can be misleading when dealing with imbalanced datasets where one class significantly outweighs the other (Swets, 1988). It is calculated as such:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 3.3

The F1 score is the harmonic mean of sensitivity and specificity and is particularly useful when the costs of false positives and false negatives are high or similar. It is a more robust measure than accuracy, especially when dealing with datasets that have an unequal class distribution (Rijsbergen, 1979). It is calculated as such:

$$F1\ Score = \frac{2(Sensitivity \times Specificity)}{Sensitivity + Specificity}$$

Equation 3.4

Each of these metrics captures different aspects of a model's performance. In sports analytics, for example, sensitivity might reflect the model's ability to accurately predict wins for a team with high accuracy, specificity could demonstrate the model's capacity to anticipate losses correctly, accuracy would provide an overall effectiveness rate of the model,

and the F1 score would balance the precision and recall of the model's predictions on the outcomes of the games (*Baumer, Jensen & Matthews, 2015)*.

3.4.2 ROC – AUC

The receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier by varying the discrimination threshold.

In the context of classification, the ROC curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1 - specificity) at various threshold settings. It provides a tool to select possibly optimal models and discard suboptimal ones.

The area under the curve (AUC) summarises the entire ROC curve in a single value, offering an aggregated measure of a model's performance across all classification thresholds. An AUC of 1.0 represents a perfect model; an AUC of 0.5 suggests a performance no better than random chance (Fawcett, 2006).

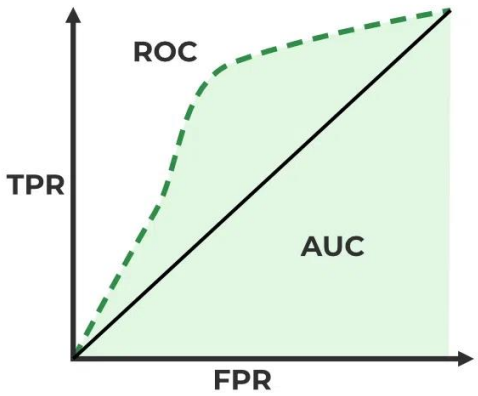

Figure 3 - ROC Curve Diagram (Image courtesy: geeksforgeeks: auc-roc-curve)

In sports analytics, ROC-AUC can be used to evaluate predictive models for outcomes like winning a match. The ROC curve can help identify the optimal probability threshold for predicting a win based on player statistics, game location, and other relevant factors (Polo & Miot, 2020).

3.4 Bernoulli Trial

A Bernoulli trial is a random experiment where there are only two possible outcomes: success or failure. Named after Swiss mathematician Jacob Bernoulli, the Bernoulli trial is the simplest non-trivial random experiment in statistical theory. The mathematical properties of

Bernoulli trials underlie the binomial distribution, which describes the number of successes in a fixed number of independent Bernoulli trials.

The significance of Bernoulli trials lies in their broad applicability across various fields. They form the foundation of the theory of probability and are particularly prominent in studies that require binary outcomes, such as clinical trials, quality control, and even simplistic models of market fluctuations.

In essence, the outcome of a Bernoulli trial with a success probability p is modelled as a Bernoulli random variable X, which takes value 1 with probability p and value 0 with probability $1 - p$. This simple framework has been instrumental in advancing probability theory and statistics, and it serves as a building block for more complex probabilistic models.

3.5 Monte Carlo Simulations

Monte Carlo simulations are computational algorithms that rely on repeated random sampling to obtain numerical results, typically one that might be deterministic but complex to solve (Metropolis & Ulam, 1949). The term “Monte Carlo” was coined by physicists working on nuclear weapons projects in the 1940s, referring to the Monte Carlo Casino in Monaco where chance and randomness are integral to the outcome of games of chance.

In a Monte Carlo simulation, the random variable of interest is simulated many times under a probabilistic model. The collection of results yields approximations of mean values, standard deviations, and percentile outcomes. It is particularly useful for assessing risk, optimising complex systems, and understanding the impact of uncertainty.

The power of Monte Carlo simulations lies in their flexibility and broad applicability; they can be adapted to any system with inherent uncertainty or complex interactions that cannot be accurately described with closed-form analytics. This methodology has become an essential tool in both theoretical and applied statistics (*Harrison, 2010)*.

3.6 Process

Integrating logistic regression, Bernoulli trials, and Monte Carlo simulation forms a robust method for predicting NBA game outcomes, incorporating uncertainty and variability inherent in sports. Logistic regression uses historical data to predict the probability of a team winning based on variables like player performance and team dynamics. These probabilities then inform Bernoulli trials, simulating each game's outcome as either a win or loss. Monte

Carlo simulation enhances this by repeating the process numerous times, generating a distribution of outcomes that reflects the potential variability and risk of game results. This combined approach not only forecasts who might win or lose but also provides insights into the likelihood and variability of these outcomes, offering a nuanced view ideal for informed decision-making in sports analytics (Mehta, Patel & Senchaudhuri, 2000).

4 Data Pre-processing

4.1 Data Cleaning

The introduction of big data in sports analytics has revolutionised the way teams, analysts, and fans understand sports and in particular basketball. Our project leverages this evolution by integrating diverse and comprehensive datasets to analyse NBA games, focusing on predicting game outcomes post the 2021 All-Star break. The methodology encompasses data verification and cleaning.

The foundation of our analysis is built on the integration of three key datasets; historical box scores of NBA games from 1949 to 2020, the 2020-2021 season data, and FiveThirtyEight's NBA Elo dataset. The two datasets of game box scores came from Kaggle by Rafael Greca. These datasets were created from a web scraper using python scraping data from basketball reference. The NBA Elo dataset by the official FiveThirtyEight website was found on github, which provided NBA Elo data from the 1946/47 season until the 2022/23 season. To validate these datasets 25 games were selected at random and examined comparing the data from the datasets against the official box scores found on the official NBA websites. From this there was no evidence of inaccurate information. The integration of these datasets aimed to create a comprehensive view of the NBA landscape over the past four decades, capturing the nuances of team dynamics.

Initial cleaning of the datasets involved standardising column names for compatibility across all datasets to ensure consistency. Variables such as game dates (‘gmDate’), team abbreviations (‘teamAbbr’, ‘opptAbbr’), and team statistics (‘team3PM’, ‘oppt3PM’, etc.) were the names given in the datasets from Kaggle, where the ‘team’ variables were from the perspective of the away teams, however, the Elo dataset had differing names for identifying teams where, generally, the away team was listed as ‘team1’ and the home team ‘team2’. So, using Microsoft Excel the variable names were manually renamed to begin with either ‘team’ or ‘oppt’ to ensure consistency across the datasets which was crucial for seamless dataset

merging and accurate data manipulation in subsequent stages. Also, given the historical depth of our data, we addressed the challenge of franchise relocations and name changes by meticulously mapping old identifiers to their current equivalents, for example the Seattle Supersonics, 'SEA', to the Oklahoma City Thunder, 'OKC', ensuring continuity in analysing team history.

Another critical aspect of our data pre-processing was setting an appropriate timeframe of the dataset. Acknowledging the significance of the 1983/84 season as the era when full comprehensive games statistics, such as blocks and steals, started being recorded, we filtered our datasets to begin from this pivotal moment in NBA analytics. Furthermore, for our purposes of predictive modelling with contemporary analyses, we excluded games beyond the 2021 All-Star break, establishing a focused dataset for model training and validation.

Also, playoff games were deliberately omitted to concentrate on the regular season which is regarded amongst players and analysts alike as a different competitive environment. This decision was underpinned by the idea that regular-season dynamics offer a unique analytical framework, differing from the high-stakes nature of playoff NBA basketball.

The creation of a unique game identification, 'game_id', for each match was an approach adopted to merge and reconcile data points across our datasets. The 'game_id' was crafted by combining the game date with the home team abbreviations, yielding a unique identifier for each game that served as a key in our data merging efforts.

This identifier became particularly useful when we discovered discrepancies between datasets. Instances where the game identifications did not match across datasets flagged potential data collection errors or inconsistencies.

Box Score Dataset Exclusive Game IDs	*Elo Ratings Dataset Exclusive Game IDs*
19921106HOU	*19921106OKC*
19941105POR	*19941105LAC*
19991105SAC	*19991105MIN*
19991106MIN	*19991106SAC*
20031030OKC	*20031030LAC*
20031031LAC	*20031031OKC*
20071219ATL	*20080307ATL*
20110304BRK	*20110304TOR*
20130117DET	*20130117NYK*
20150115MIL	*20150115NYK*
20200815POR	

Table 1 - Exclusive 'game_id' in each dataset

Table 1 details all the games which were unique to each respective dataset. To address these, we initiated a manual verification process, consulting official NBA game records to determine the accuracy of each entry. This process was meticulous, with each discrepancy manually checked against information from Basketball Reference. Games that were not aligned were review to identify the root cause of the mismatch. On examination it was determined that the cause of these discrepancies was due to these games being played in neutral venues. However, due to this these games were excluded from our dataset since our potential models may use team location as a predictor. Upon completion of this process, we were still left with the following discrepancies: '20071219ATL', '20080307ATL' and '20200815POR'. First looking at 20200815POR, this was a play-in game between the Portland Trailblazers and Memphis Grizzlies that was wrongly recorded as a regular season game in our box score dataset, which was then ultimately excluded from the dataset.

However, the last two remaining Game IDS are representative of a game between the Miami Heat and Atlanta Hawks that was a unique case. Initially this game was played on the 19th December 2007, however, there was an error where one of the player, Shaquille O'Neal, was wrongfully ejected late in the game due to a bookkeeping error, where he was wrongly awarded an additional foul. The game was then appealed, and the final 52 seconds were set to be replayed before their rematch later that season, this is detailed in a The Ringer magazine article. However, due to the bizarre nature of this game we decided to exclude this game from our datasets.

Now, the datasets were combined using the unique game identifier to ensure the Elo and game statistics remained consistent and gave a comprehensive view of each game. Also,

another aspect of our data pre-processing was the switching of perspectives for each game. We recognised the importance of analysing games from both the home and away team viewpoints to capture the full spectrum of game dynamics. For this, using programming in R we switched the 'team' and 'oppt' prefixes for all variables, effectively doubling our dataset while maintaining its integrity. This perspective switch allowed us to mirror each game, ensuring that our analysis encompasses all potential location advantages.

With our cleaned dataset the next step involved expanding our data for the purposes of prediction through feature engineering. We introduced variables such as moving averages of team results and team statistics for each individual season. The mean results would be used for hypothesising their influence on game outcomes in our predictive models.

4.2 Feature Engineering

Feature engineering is a pivotal step in predictive modelling, where raw data are transformed into a format that better represents the underlying problem for predictive models. For this project, we have concentrated on calculating moving averages for key statistical measures for both NBA teams and their opponents to be utilised for prediction. These measures include assists, 3-pointers made (3PM), 2-pointers made (2PM), turnovers (TO), blocks (BLK), steals (STL), personal fouls (PF), free throws made (FTM), total rebounds (TRB), and defensive ratings (DRTG). Moving averages are used to smooth out short-term fluctuations and highlight longer-term trends, making them invaluable for assessing team performance stability.

The ‘rollapply’ function from the ‘zoo’ package in R is instrumental in calculating these moving averages. Applied to a rolling window of 10 games, ‘rollapply’ computes the mean for each statistical category, ensuring that each calculation includes data from the 10 games preceding it. This approach provides a smoothed performance metric, mitigating the impact of outliers and capturing the team's performance trend.

To prevent lookahead bias, which could invalidate our predictive model by using future information, we apply the lag function to the computed averages. This adjustment ensures that the model predictions are based on past data only, adhering to realistic forecasting scenarios. By integrating these engineered features, our dataset now reflects more nuanced aspects of team and opponent performances. This enhancement is anticipated to

significantly improve the predictive accuracy of our models, allowing for more sophisticated analysis and insights into the factors driving NBA game outcomes.

There we now have determined the following predictors:

Predictor	*Description*
teamLoc	*Indicator of whether team is home or away*
teamForm	*Team win rate in past 10 games*
opptForm	*Opponent win rate in past 10 games*
teamAvgAST	*Team average assists in past 10 games*
teamAvg3PM	*Team average 3 pointers made in past 10 games*
teamAvg2PM	*Team average 2 pointers made in past 10 games*
TeamAvgFTM	*Team average free throws made in past 10 games*
teamAvgTO	*Team average turnovers in past 10 games*
teamAvgBLK	*Team average blocks in past 10 games*
teamAvgSTL	*Team average steals in past 10 games*
teamAvgPF	*Team average personal fouls in past 10 games*
teamAvgTRB	*Team average total rebounds in past 10 games*
opptAvgDRTG	*Opponent average defensive rating in past 10 games*

Table 2 - Engineered Predictors for Model Prediction

Also, weights were assigned to the data from different years to reflect their relevance to current trends. The most recent games in 2021 are weighted highest, while older games receive progressively lower weights, acknowledging their diminishing influence on present-day outcomes. This will be examined why in the exploratory data analysis.

Furthermore, the dataset divided a training dataset and a test dataset. The split was based on the All-Star break. The training dataset contains all games that occurred before the All-Star while the test dataset comprises games played after.

5 Prediction

5.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical initial step in the data analysis process, involving the examination of our dataset to discover patterns, spot anomalies, test a hypothesis, or check assumptions. It provides important insights which can have direct effect on analysis and modelling.

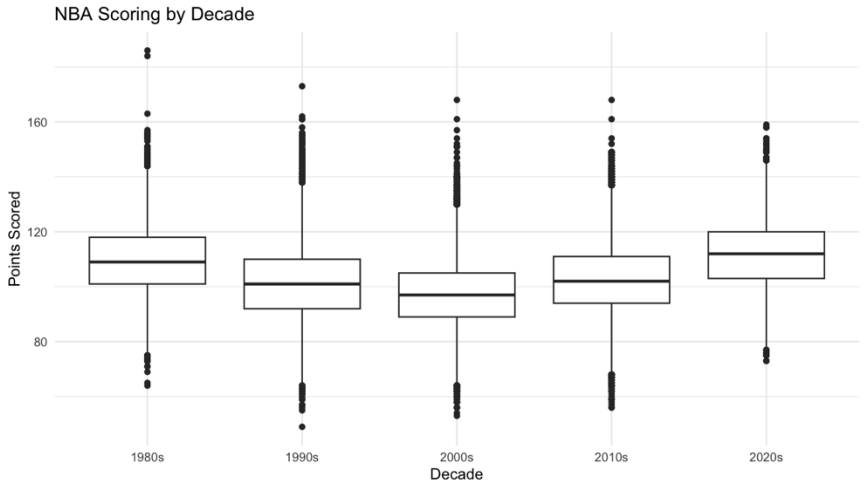

Figure 4 - NBA points scoring by decade

The boxplot displaying NBA scoring by decade provides a visual depiction of the distribution of points scored across the decades in the NBA. The 1980s show a relatively high median, indicating a period of high-scoring games. As we progress through the decades, we notice a slight decrease in median scores, reflecting a potential shift towards defensive strategies or changes in game pace. Until we approach the 2010s where the median starts to increase again. This boxplot gives evidence to the ever-evolving tactics of NBA basketball and emphasises that weighting may need to be applied into our predictive models.

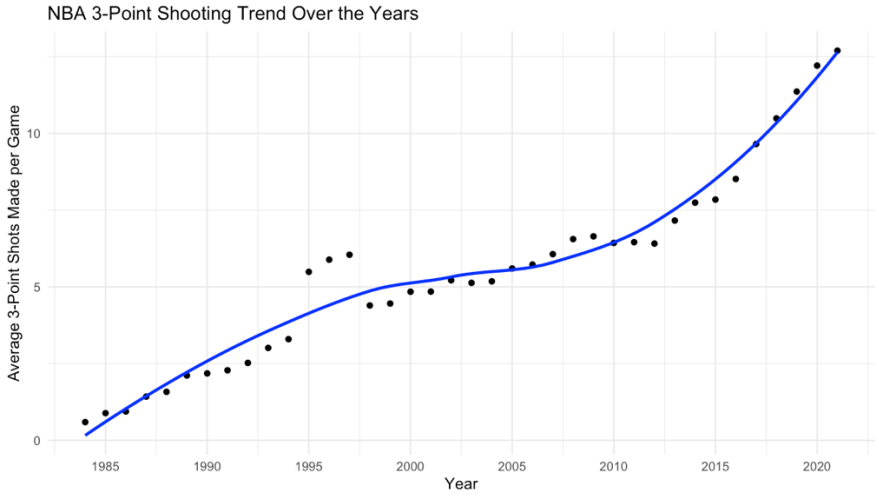

Figure 5 - NBA 3 Point Shooting Since 1983

The time series plot highlights the evolution of the 3-point shot since 1983. The data points, fitted with a curve, show a gradual increase in the average 3-point shots made per game, with a significant increase in more recent years further highlighting the modern efficiency of

the three-point shot. On this plot there are outliers on our curve from 1985 until 1987, this was due to a reduction in distance for the three-point line in those years.

Distributions of potential predictors were assessed to investigate potential correlation with game results. From this, we did not encounter any anomalies or unusual findings in these predictors and all assessed variables had correlations with results as would be expected, apart from rebounding.

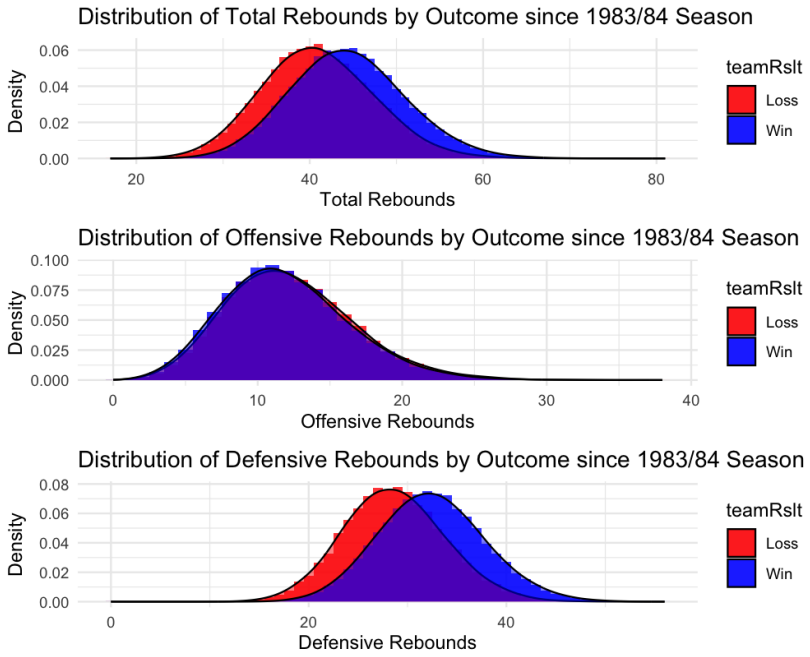

Figure 6 - Distribution of Rebounds by outcome

The distribution of total rebounds, offensive rebounds, and defensive rebounds by game outcome reveals interesting insights. While offensive rebounds do not seem to be a strong indicator of winning games, the total and defensive rebounds show a more distinct distribution between winning and losing outcomes. Wins are associated with higher total and defensive rebound numbers, suggesting that a team's ability to secure rebounds on the defensive end contributes significantly to their chances of winning. Therefore, it was decided to only include total rebounds as a predictor in our models rather than focusing on offensive rebounds.

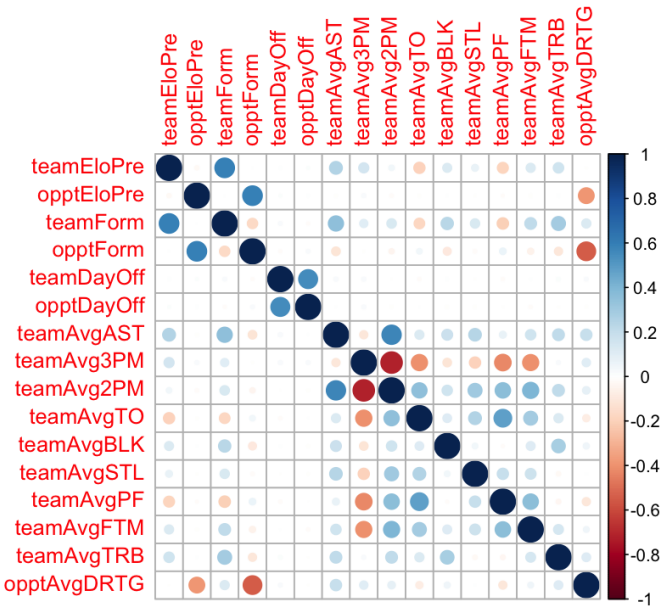

Figure 7 - Correlation Matrix of potential predictors

The correlation matrix plot helps us examine the potential multicollinearity between predictors, which may violate the assumptions of our models. Variables that are highly correlated with one another may not both be necessary in our models, as they could contribute to multicollinearity issues. From the matrix, we can dissect which variables correlate with each other and may need to be evaluated for exclusion or transformation to ensure the stability and interpretability of our predictive models. From this, we can see a strong negative correlation between average 2-point shots and average 3-point shots, this makes sense as if a team is shooting more 2 pointers, they will have less opportunity to shoot 3 pointers and vice versa. In constructing our models, we'll consider the following predictors: the team of interests Elo rating before the game; the opposing team's Elo rating before the game; the location of the team's game (home or away); the recent performance trend of the team; the recent performance trend of the opposing team; and various average statistics for the team including assists, three-point field goals made, two-point field goals made, free throws made, turnovers, blocks, steals, personal fouls, total rebounds, and defensive rating. These will be integral in the predictive modelling of game outcomes.

5.2 Model Building

Our first comprehensive logistic regression model was established to predict NBA game outcomes based on factors determined in our exploratory data analysis. This model was built using our training dataset.

The table below showcases the model's coefficients, which reflect the influence of each predictor variable on the likelihood of the team of interest winning:

Variable Name	*Coefficient*	*p-value*
(Intercept)	*-0.97*	*0.0014*
teamEloPre	*0.0053*	*<2e-16*
opptEloPre	*-0.0055*	*<2e-16*
teamLoc	*0.92*	*<2e-16*
teamDayOff	*0.013*	*0.012*
opptDayOff	*-0.013*	*0.023*
teamForm	*0.11*	*0.012*
opptForm	*-1.52*	*0.00034*
teamAvgAST	*0.018*	*2.5e-07*
teamAvg3PM	*-0.0079*	*0.077*
teamAvg2PM	*-0.013*	*9.1e-05*
teamAvgFTM	*0.0088*	*0.000993*
teamAvgTO	*-0.014*	*0.00066*
teamAvgBLK	*0.020*	*0.00052*
teamAvgSTL	*0.017*	*0.0018*
teamAvgPF	*0.0024*	*0.48*
teamAvgTRB	*0.0055*	*0.042*
teamAvgDRTG	*0.0027*	*0.12*

Table 3 - Initial Logistic Regression Model

Key takeaways from figure 4 include the strong influence of a team's Elo rating before the game and the benefit of playing at home. Additionally, team statistics like assists, blocks, and steals are positively linked to winning. However, the model also presents some unexpected findings, such as the negative impact of points made from two and three-point field goals, which may be due to the multicollinearity identified in the exploratory analysis.

To refine the model, simplification and exploring potential interactions between variables were examined. By focusing on these interactions and simplifications, the model can become more interpretable and potentially more accurate in predicting game outcomes. Through this process a final model was produced:

Variable Name	*Coefficient*	*p-value*
(Intercept)	*-1.1*	*0.00012*
teamEloPre	*0.0055*	*<2e-16*
opptEloPre	*-0.0057*	*<2e-16*
teamLoc	*0.98*	*<2e-16*
teamAvgAST	*0.0098*	*0.00016*
teamAvg3PM	*0.0059*	*0.035*
teamAvgFTM	*0.0095*	*0.00020*
teamAvgTO	*-0.014*	*0.00050*
teamAvgBLK	*0.024*	*1.83e-05*
teamAvgSTL	*0.015*	*0.0043*
teamAvgDRTG	*0.0035*	*0.030*

Table 4 - Final Logisitc Regression Model

This refined logistic regression model for NBA game predictions highlights several key predictors of winning. Once again, a higher team pre-game Elo rating correlates with increased chances of victory, emphasising the importance of a team's strength. Home advantage is also significant, with teams more likely to win at their own venue.

Offensively, assists and three-pointers made both positively affect winning probabilities, indicating the value of teamwork and effective scoring. Free-throw accuracy also contributes positively, whereas turnovers negatively impact the chances of winning.

Defensively, blocks and steals are beneficial, underlining the importance of a solid defensive game. Overall, this model presents a detailed view of the factors influencing NBA game outcomes.

5.3 Model Validation

Now from our final model we want to assess the model's predictive power and its capability to accurately classify games into wins and losses. This evaluation was performed on both the training and test datasets to ensure the model's robustness.

We initiated our analysis by generating win probabilities for each game using the logistic regression model. These probabilities were derived using the ‘predict’ function in r, to obtain outcomes in terms of win likelihood. A threshold of 0.5 was applied to these predicted probabilities to classify each game's outcome, where win likelihoods equal to or greater than 0.5 were assigned a win and win likelihoods below 0.5 were assigned a loss. This binary classification approach is standard practice in logistic regression models dealing with two possible outcomes.

The cornerstone of our performance evaluation was the confusion matrix, which provided a detailed view of the model's predictions against the actual outcomes. From the confusion matrix, we calculated the sensitivity and specificity of our model. The model exhibited a sensitivity of 68.28% and a specificity of 68.33%, demonstrating its balanced capability to correctly identify both wins and losses.

The overall accuracy of the model was calculated to be 68.25%, indicating that it correctly predicted the outcome of approximately 68.25% of the games in the training dataset. Additionally, the F1 score, a harmonic mean of precision and sensitivity, was computed to be 0.6831. This metric provided further confirmation of the model's balanced performance between sensitivity and specificity.

Now looking at our ROC curve:

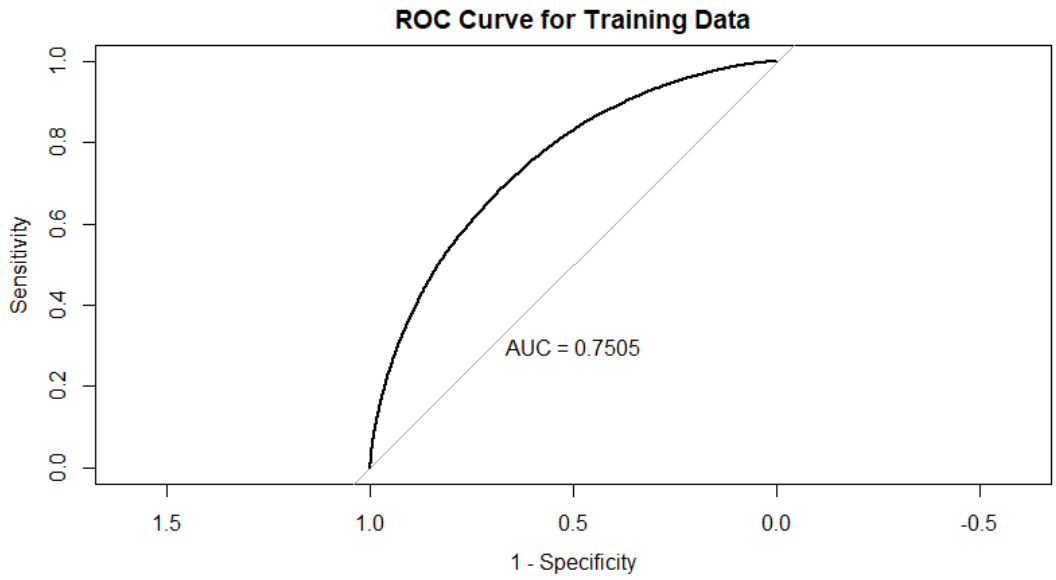

Figure 8 - ROC Curve on training data

The ROC curve itself, being well above the diagonal line, shows that the model performs significantly better than a random classifier on the training dataset. The AUC of 0.7505 indicates that the logistic regression model has a good level of discrimination. This level of AUC indicates that the model has good predictive power and can be a reliable tool for forecasting outcomes in NBA games.

Now we want to see the models, predictive power on unseen data. Analysing our logistic regression model's performance on unseen data, we conducted an analysis using our test dataset, which has been segregated from the training dataset to provide an unbiased evaluation of the model's predictive power.

We began by applying our logistic regression model to the test data to predict outcomes of NBA games, assigning predicted probabilities of wins and transforming these into binary predictions based on a 0.5 threshold. We then once again, constructed a confusion matrix to compare these predictions against the actual results.

Our model achieved a sensitivity rate of 64.08%, the specificity rate was 61.97%, showing the overall accuracy of the model on the test data stood at 64.90%, a respectable rate for out-of-sample predictions and the F1 score was calculated to be 0.63.

To further validating the model's predictive capability on unseen data the ROC curve for the test data was plotted:

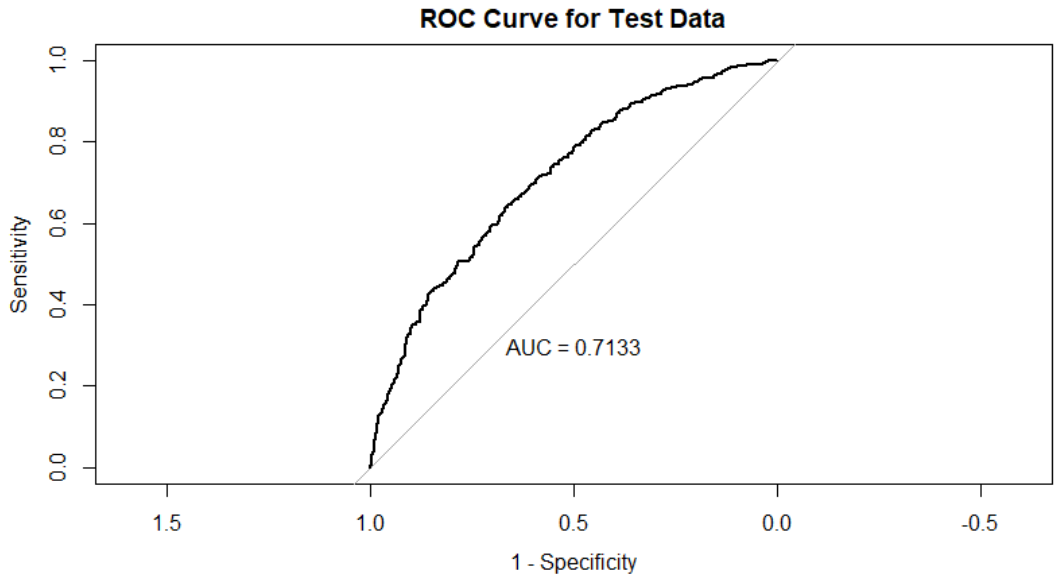

Figure 9 - ROC curve on test data

From our curve we obtained an AUC value of 0.7133 which suggests that the model has good discriminatory ability, although there is room for improvement.

In conclusion, our model demonstrates a good level of predictive accuracy, sensitivity, and specificity when applied to new data, as evidenced by the AUC from the ROC curve analysis. These results show that the model holds practical predictive value, reinforcing its applicability in forecasting NBA game outcomes.

5.4 Three Point Analysis

Now we want to further explore how NBA 3 pointers has become increasingly important as evidenced in our exploratory data analysis.

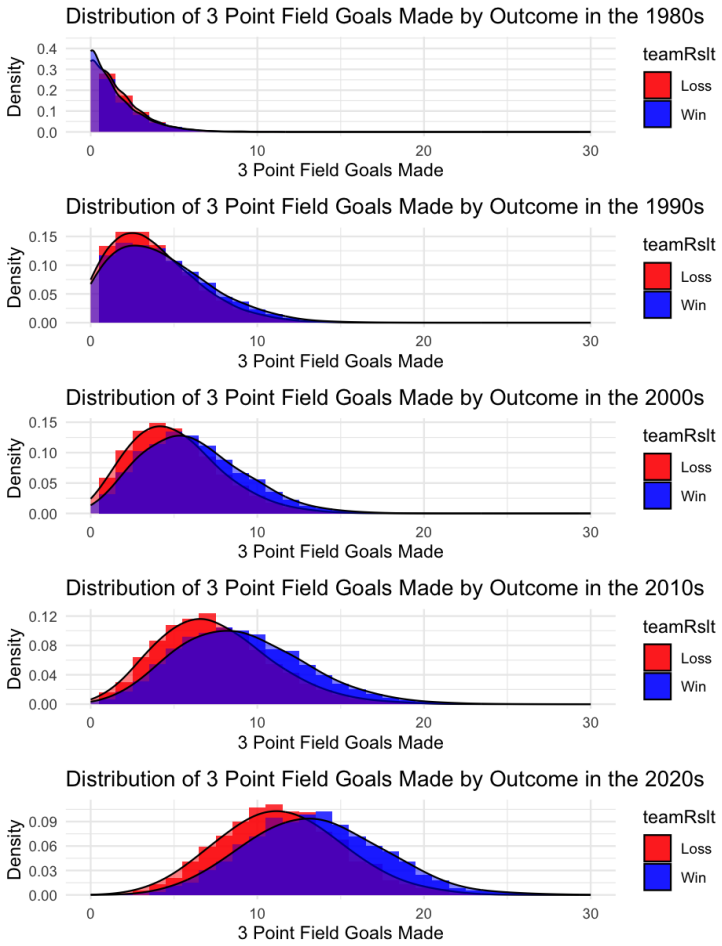

Figure 10 - Distribution of 3 Point FGM over the decades

Figure 10 provides a series of density plots for the distribution of 3-point field goals made by outcome across different decades. From this we can see indicates a significant evolution in basketball strategy, particularly in the reliance on 3-point shots. From the 1980s to the 2020s, there's a notable trend showing an increase in both the number of 3-point shots made and their impact on game outcomes.

In the 1980s, when the 3-point shot was introduced, the density plots reveal a relatively low frequency of games with high 3-point shots made, and the distributions for wins and losses are closely aligned, suggesting that 3-point shots were not a major factor in winning games.

As we progress into the 1990s, the trend continues. The plots for the decade show only a slight increase in the number of 3-point shots made by winning teams, implying that the 3-point shot still wasn't considered essential for securing wins. The slight shift to the right for winning teams is visible, yet it's clear that the high-volume 3-point shooting strategy had not become a dominant strategy.

The 2010s mark a noticeable change. The win and loss distributions now show a clear gap, with winning teams making a higher number of 3-point shots. This indicates a significant growing recognition of the value of the 3-point shot in winning games.

By the 2020s, the strategy has fully evolved. There's a marked separation between the win and loss plots, with the peak for wins substantially higher. This reflects a tactical shift where 3-point shooting has become integral to offensive play. This evolution has significant implications for team dynamics, coaching strategies, and the evaluation of player skills. The increasing importance of 3-point proficiency may lead teams to prioritise sharpshooters and plays that open the perimeter.

For predictive modelling, these shifts emphasise the need for continual adaptation. The changes in the plots over the decades underline the need for our predictive models to adapt over time, as the factors that influence game outcomes evidently shift.

5.5 Model Simulation

Our objective was to simulate and understand the progression of NBA team standings throughout the season, after the all-star break, by utilising a predictive model that incorporates Elo ratings. Elo ratings are dynamic indicators of team performance. These ratings fluctuate based on game outcomes.

The simulation leverages a logistic regression model that factors in these Elo ratings, along with other variables indicative of team performance, to forecast the probability of a team winning a specific game. Since Elo ratings are dependent on outcomes, differentiating from the previous predictions where Elo ratings were already known for future simulations to replicate real life, we will need to ensure they are updated after each game to accurately reflect the latest team performance levels.

We begin with the initial Elo ratings in our test dataset to establish the current standing of each team. As we progress through the simulation, each game's outcome is predicted using

the logistic regression model, which considers the existing Elo ratings and other predictive indicators. The outcome for each game is then simulated using a Bernoulli trial, reflecting the binary nature of win-loss results in sports.

The Elo ratings are updated following each simulated game, with the winning team's rating going up and the losing team's going down in accordance with the established Elo calculation method. However, due to the limitations of what we are simulating, we are unable to simulate margin of victory. So, for this we will not be incorporating the G factor established in FiveThirtyEight's calculation for Elo rating. This updating is iterative, ensuring that the predictions for forthcoming games are made using the most recent Elo ratings, much like the continual adjustments that would occur during an actual NBA season.

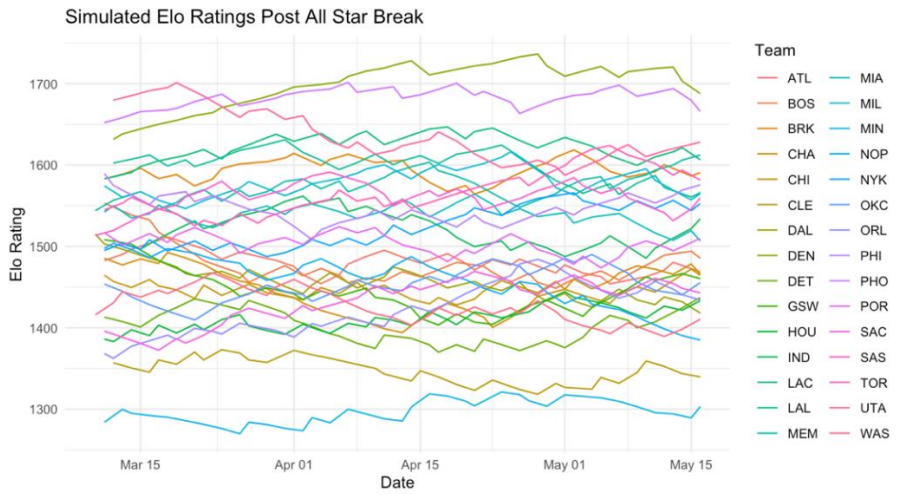

Figure 11 - Simulated Elo Ratings Post All Star Break

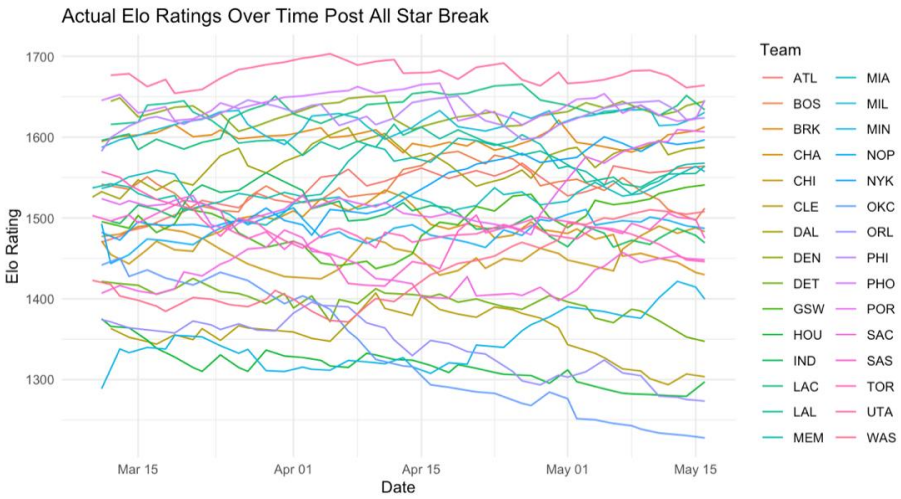

Figure 12 - Actual Elo Ratings Post All Star Break

Figures 11 and 12 present a compelling visual comparison between simulated and actual Elo ratings of NBA teams after the All-Star break, respectively. These plots illustrate the trajectories of teams' strengths as estimated by their Elo ratings.

In Figure 11, the Elo ratings are the result of a series of simulated outcomes based on our logistic regression model. This model considers not only the current Elo ratings but also other performance-related variables, predicting the likelihood of a team winning each game. These predictions inform an iterative process where game outcomes are simulated using a Bernoulli trial, and subsequent Elo ratings are updated accordingly. This simulation process reflects a potential path that teams' performances could take, strictly based on the probabilistic model.

Figure 12, on the other hand, displays the actual Elo ratings that evolved as the season progressed. This provides a factual account of how teams' performances varied over the same period, serving as a benchmark against which the simulated outcomes can be evaluated.

Comparing the two figures allows for an analysis of how closely the simulated data aligns with reality. Discrepancies between the two could highlight the limitations of the predictive model or indicate unexpected turns in the season's events that the model could not account for. As we can see, the simulation is not perfect but there is nothing too glaringly concerning so we will continue with the simulation.

The simulation showcased in Figure 11 represents just one iteration of a possible future, based on the Elo ratings system embedded within our logistic regression model. A single simulation can be influenced by random variance and might not capture the full spectrum of potential outcomes; therefore, we extended our analysis through Monte Carlo simulations. This approach involves running multiple simulations, aggregating the results to obtain a more comprehensive and statistically robust forecast.

By conducting a multitude of simulations, in this case 1000, each representing a different possible future, we can average out the randomness inherent in any single simulation. This method allows us to converge on a more stable and accurate prediction of team standings. The variability in outcomes due to the inherent uncertainty in sporting events is thereby reduced, leading to a projection that is less dependent on the randomness of one individual simulations.

6 Results

First, the results from our simulation in figure 11:

Actual Pos.	Team	Actual Wins	Actual Losses	Sim Total Wins	Sim Total Losses	Sim Pos.
1	PHI	49	23	44	28	2
2	BRK	48	24	46	26	1
3	MIL	46	26	43	29	3
4	ATL	41	31	33	39	9
5	NYK	41	31	40	32	4
6	MIA	40	32	37	35	7
7	BOS	36	36	37	35	8
8	IND	34	38	34	38	6
9	WAS	34	38	29	43	13
10	CHA	33	39	32	40	10
11	CHI	31	41	31	41	11
12	TOR	27	45	39	33	5
13	CLE	22	50	24	48	14
14	ORL	21	51	29	43	12
15	DET	20	52	24	48	15

Table 5 – Eastern Conference Simulated Standings for first iteration

Actual Pos.	Team	Actual Wins	Actual Losses	Sim Total Wins	Sim Total Losses	Sim Pos.
1	UTA	52	20	51	21	2
2	PHO	51	21	53	19	1
3	DEN	47	25	51	21	3
4	LAC	47	25	42	30	6
5	DAL	42	30	43	29	4
6	LAL	42	30	39	33	8
7	POR	42	30	36	36	9
8	GSW	39	33	31	41	11
9	MEM	38	34	43	29	5
10	SAS	33	39	42	30	7
11	NOP	31	41	29	43	12
12	SAC	31	41	34	38	10
13	MIN	23	49	15	57	15
14	OKC	22	50	26	46	13
15	HOU	17	55	25	47	14

Table 6- Western Conference Simulated Standings for first iteration

Table 6 for the Eastern Conference and Table 7 for the Western Conference, we witness some intriguing differences between the actual season outcomes and the simulated results post the All-Star break. The Philadelphia 76ers, while topping the actual Eastern standings, are projected second in the simulated scenario, with the Brooklyn Nets, second, taking the top simulated position. This flip suggests a close contest between the two for the conference's dominance. Other notable differences include the Atlanta Hawks and the New York Knicks, whose strong actual performances seem underrated by the simulation, suggesting a potential underestimation of their capabilities by the model.

In the Western Conference, the simulation closely mirrors the actual top standings, with the Utah Jazz and Phoenix Suns almost accurately positioned. However, the Los Angeles Clippers and the Los Angeles Lakers fall short in the simulated standings compared to their actual achievements, hinting at potential overachievement or the simulation's conservative estimate of star-driven teams.

These simulations, while insightful, represent only a single iteration. To obtain a more comprehensive understanding and mitigate the impact of variance, we have proceeded with aggregated Monte Carlo simulations. These simulations will offer a more robust picture by averaging results across multiple iterations, providing a statistically grounded forecast of the standings.

Pos.	Team	
1	PHI	PHI
2	BRK	BRK
3	MIL	MIL
4	ATL	BOS
5	NYK	MIA
6	MIA	TOR
7	BOS	NYK
8	IND	IND
9	WAS	CHA
10	CHA	ATL
11	CHI	CHI
12	TOR	WAS
13	CLE	DET
14	ORL	CLE
15	DET	ORL

Table 7 - Eastern Conference Simulation

Pos.	Team	
1	UTA	UTA
2	PHO	PHO
3	DEN	DEN
4	LAC	LAL
5	DAL	LAC
6	LAL	POR
7	POR	DAL
8	GSW	MEM
9	MEM	SAS
10	SAS	GSW
11	NOP	NOP
12	SAC	OKC
13	MIN	SAC
14	OKC	HOU
15	HOU	MIN

Table 8 - Western Conference Simulation

Table 7 and Table 8 reveal interesting patterns in predictive accuracy, where the left column of team names represents actual rankings and the right indicating the simulated ranking. For both tables, teams with rankings that match exactly between actual outcomes and simulations are highlighted in green. These include top performers like the Philadelphia 76ers and Brooklyn Nets in the East, and Utah Jazz and Phoenix Suns in the West, indicating a high level of predictability at the top of the standings.

Teams with a discrepancy of one or two positions are marked in orange. This slight variance suggests that while the simulations closely capture team performances, minor fluctuations, likely due to the unpredictable nature of sports competitions, can lead to small ranking shifts. Examples include New York Knicks and Miami Heat in the East, and Los Angeles Clippers and Portland Trailblazers in the West, reflecting the challenges in forecasting middle-tier team standings with absolute precision.

Red highlights indicate teams whose simulated and actual rankings diverge by three or more positions, signifying areas where the model's predictive capabilities face limitations. Many cases are seen in the Eastern Conference, underscoring the potential impact of unforeseen factors like injuries, trades, and in-season improvements or declines on team performance.

Overall, Tables 7 and 8 offer a nuanced view of the simulation's effectiveness. While the model demonstrates a robust ability to predict the top-tier teams' performances accurately, the variability increases for middle and lower-tier teams.

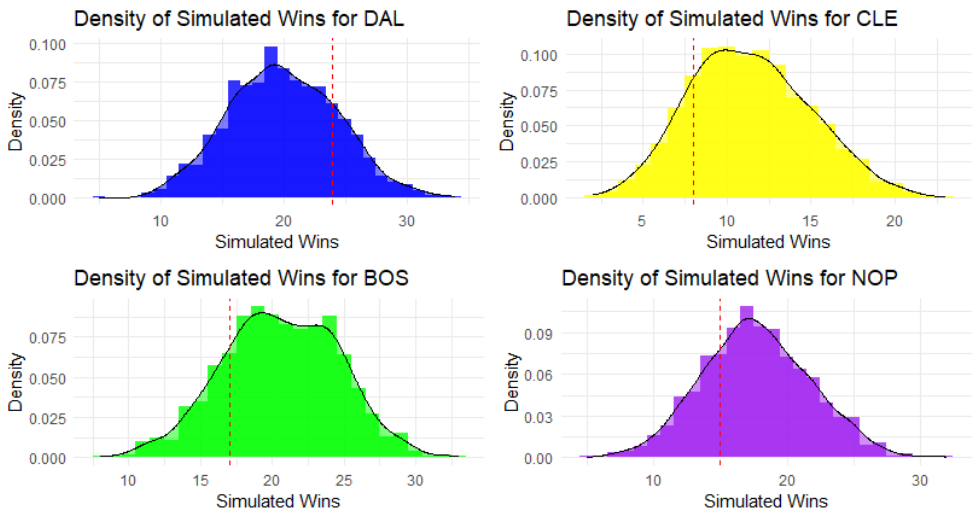

Figure 13 - Density of Simulated Wins

Figure 13 displays the density for the simulated wins of certain NBA teams which visualises the variation in performance across the simulations. Dallas and Boston show distributions centred around the mid-20s. Cleveland and New Orleans, on the other hand, display peaks at lower win totals, suggesting fewer wins on average over the simulation period. The spread of each distribution indicates the variability in the simulation outcomes, with some teams showing a wider range of possible wins. The red dashed lines mark the actual number of wins each team had during the simulation period, and their position relative to the peak of each distribution provides a visual cue on the simulation's central tendency compared to the real-world results. Overall, these plots collectively offer a snapshot of the expected performance range for each team given the model's assumptions and the data used.

7 Discussion

Our investigation into the predictive power of logistic regression models, Elo ratings, and Monte Carlo simulations on NBA game outcomes has yielded intriguing insights. This project emphasises the evolving nature of basketball, particularly through the lens of the three-point shot, and its growing influence on game strategies and outcomes.

Our findings reveal that logistic regression models, when combined with Elo ratings, offer a robust framework for predicting NBA game outcomes. The incorporation of the three-point shot as a predictive variable not only reflects its strategic importance in modern basketball but also enhances the model's accuracy. This aligns with the work of Freitas (2021), who highlighted the three-point shot's ascendancy in basketball strategy.

While our models provide a robust tool for prediction, they are not without limitations. The dynamic nature of sports, including unpredictable elements like player injuries, mid-season trades, and the psychological aspects of the game, remains challenging to encapsulate fully in any statistical model. Additionally, the reliance on historical data may not fully account for the rapid strategic evolutions or the emergence of transformative players.

Future research could explore the integration of real-time player performance metrics, the impact of coaching strategies, and the role of player psychology in game outcomes. Furthermore, expanding the dataset to include more granular details, such as player positioning and in-game events, could unveil deeper insights into the subtleties that influence game dynamics.

8 Conclusion

In conclusion, this project has shed light on the predictive capabilities of combining logistic regression with Elo ratings and Monte Carlo simulations in forecasting NBA game outcomes. By highlighting the strategic importance of the three-point shot, our research contributes to the ongoing dialogue about analytics in basketball, offering a model that balances historical insights with the nuances of modern strategy. As the NBA continues to evolve, so too will the analytical frameworks that seek to understand its complexities.

9 Appendix

The code used for feature engineering:

```
data <- data %>%
group_by(teamAbbr, year) %>%
mutate(
teamAvgAST = lag(rollapply(teamAST, width = 10, FUN = mean, partial = TRUE, align = 'right')),
teamAvg3PM = lag(rollapply(team3PM, width = 10, FUN = mean, partial = TRUE, align = 'right')),
teamAvg2PM = lag(rollapply(team2PM, width = 10, FUN = mean, partial = TRUE, align = 'right')),
teamAvgTO = lag(rollapply(teamTO, width = 10, FUN = mean, partial = TRUE, align = 'right')),
teamAvgBLK = lag(rollapply(teamBLK, width = 10, FUN = mean, partial = TRUE, align = 'right')),
teamAvgSTL = lag(rollapply(teamSTL, width = 10, FUN = mean, partial = TRUE, align = 'right')),
teamAvgPF = lag(rollapply(teamPF, width = 10, FUN = mean, partial = TRUE, align = 'right')),
teamAvgFTM = lag(rollapply(teamFTM, width = 10, FUN = mean, partial = TRUE, align = 'right')),
teamAvgTRB = lag(rollapply(teamTRB, width = 10, FUN = mean, partial = TRUE, align = 'right'))
)
```

10 References

"Atlanta Hawks vs. Miami Heat: Replay of a Classic NBA Game from March 8, 2008." The Ringer, 2018. Available at: https://www.theringer.com/nba/2018/3/7/17088350/atlanta-hawks-miami-heat-replay-nba-game-march-8-2008 [Accessed 20 March 2024].

Cox, D.R., 1958. The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology, 20(2), pp.215-232.

Elo, A.E. and Sloan, S., 1978. The rating of chessplayers: Past and present.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27(8), pp.861-874.

Silver, N. and Fischer-Baum, R. 2015. How We Calculate NBA Elo Ratings. FiveThirtyEight. Available at: https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/ [Accessed 20 March 2024].

FiveThirtyEight. NBA Elo data. GitHub. (Last update 16 December 2022). Available at: https://github.com/fivethirtyeight/data/tree/master/nba-elo [Downloaded January 2024].

Freitas, L., 2021. Shot distribution in the NBA: did we see when 3-point shots became popular?. German Journal of Exercise and Sport Research, 51(2), pp.237-240.

GeeksforGeeks, 2024. AUC-ROC Curve. Available at: https://www.geeksforgeeks.org/auc-roc-curve/ [Accessed: March 2024].

Greca, R. NBA games box score since. Kaggle. (Last updated 26 July 2021). Available at: https://www.kaggle.com/datasets/rafaelgreca/nba-games-box-score-since-1949 [Downloaded January 2024].

Harrison, R.L., 2010, January. Introduction to monte carlo simulation. In AIP conference proceedings (Vol. 1204, p. 17). NIH Public Access.

Hosmer Jr, D.W., Lemeshow, S., and Sturdivant, R.X., 2013. Applied logistic regression. John Wiley & Sons.

Hvattum, L.M. and Arntzen, H., 2010. Using ELO ratings for match result prediction in association football. International Journal of forecasting, 26(3), pp.460-470.

Kovalchik, S., 2020. Extension of the Elo rating system to margin of victory. International Journal of Forecasting, 36(4), pp.1329-1341.

Marveldoss, R.E.D., 2018. An Elo-Based Approach to Model Team Players and Predict the Outcome of Games (Doctoral dissertation).

Mehta, C.R., Patel, N.R. and Senchaudhuri, P., 2000. Efficient Monte Carlo methods for conditional logistic regression. Journal of the American Statistical Association, 95(449), pp.99-108.

Metropolis, N. and Ulam, S., 1949. The Monte Carlo method. Journal of the American Statistical Association, 44(247), pp.335-341.

Narkhede, S. (2018). Understanding Confusion Matrix. Towards Data Science. Available at: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 [Accessed 21 March 2024].

National Basketball Association (NBA). Official Website. Available at: https://www.nba.com/ [Accessed February 2024].

Nevill, Alan & Holder, R., 1999. Home advantage in sport: an overview of studies on the advantage of playing at home. Sports Medicine (Auckland, N.Z.), 28, pp.221-36.

Polo, T.C.F. and Miot, H.A., 2020. Use of ROC curves in clinical and experimental studies. Jornal vascular brasileiro, 19, p.e20200186.

Rijsbergen, C.V., 1979. Information retrieval. Butterworth-Heinemann.

Sports Reference LLC. (2024). Basketball Reference. Available at: https://www.basketball-reference.com/ [Accessed February 2024].

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science, 240(4857), pp.1285-1293.

11 Bibliography

Dataquest, 2022. Predict NBA Games With Python And Machine Learning. Dataquest Project Walkthroughs. YouTube [video] Available at: <https://www.youtube.com/watch?v=egTylm6C2is> [Accessed January 2024].

*Horvat, T., Job, J., Logozar, R. and Livada, Č., 2023. A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games. *Symmetry*, 15(4), p.798.*

Houde, M., 2021. Predicting the Outcome of NBA Games.

Marin, P. (2023) Netty - my personal NBA game-winner predictor, Medium. Available at: https://medium.com/data-sports/netty-my-personal-nba-game-winner-predictor-63cc728025c5 [Accessed: January 2024].

Weiner, J., 2021. Predicting the outcome of NBA games with Machine Learning: How we used (and you can too) machine learning to better understand the role statistics play in sports. Towards Data Science. Available at: https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20 [Accessed: January 2024].