GHENT UNIVERSITY

FACULTY OF SCIENCES

Score Modelling of NBA and EuroLeague Games Using Compound Poisson Simulation

Emiel Platjouw

Master dissertation submitted to

obtain the degree of

Master of Statistical Data Analysis

Promotor: Prof. Dr. Christophe Ley

Department of Applied Mathematics,

Computer Science and Statistics

Academic year 2019-2020

GHENT UNIVERSITY

FACULTY OF SCIENCES

Score Modelling of NBA and EuroLeague Games Using Compound Poisson Simulation

Emiel Platjouw

Master dissertation submitted to

obtain the degree of

Master of Statistical Data Analysis

Promotor: Prof. Dr. Christophe Ley

Department of Applied Mathematics,

Computer Science and Statistics

Academic year 2019-2020

Acknowledgements

This paper and the research behind it would not have been possible without the exceptional support of my promotor, Prof. Dr. Christophe Ley. He helped me channel my enthusiasm in basketball analytics and provided me with a research project I thoroughly enjoyed. His patience, knowledge and exacting attention to detail has been an inspiration and kept me on track from my first unstructured ideas to the final draft of this paper.

Further, I would like to thank everyone involved in the organisation and teaching of the Master of Statistical Data Analysis. It was a tough journey and I spent many times being discouraged by the amount of work put in front of me but it was worth it. I can honestly say this was probably the most educationally enriching year of my academic career and will be invaluable in my further professional career.

I would also like to acknowledge the NBA and the EuroLeague for recording all the data used in this paper and making this data available to the general public for free.

Finally, I would like to thank my family and friends for the support they have given me. A special thank you goes towards my sister and my girlfriend for their help in proofreading this paper and their helpful suggestions that shaped this paper into what it is now.

Table of Contents

1 Summary

A lot of research in sports analytics is centred around identifying the factors that distinguish winning teams from losing teams. That information is used by coaches to improve their team, by the betting industry to predict outcomes, or even by regular people to discuss the upcoming game of their favourite team. This paper aims to predict the outcome of basketball games of the two most popular basketball leagues, the NBA and the EuroLeague, in the last 19 seasons.

The concept of possessions forms an integral part of contemporary advanced basketball analytics. It is, however, failing to capture valuable information useful for our purpose. A single possession can contain multiple plays and plays can thus represent the sequence of events in a basketball game more accurately than a possession. In this paper, we first try to predict the total number of plays, and the number of zero-point plays, one-point plays, two-point plays, three-point plays, and four-point plays two opposing teams will have in a game. We use penalized Poisson regression models using Elastic Net regularisation for this purpose, with box scores statistics and play-by-play data from season 2000-2001 up to and including season 2017-2018. Using these predictions, each game is simulated with a compound Poisson process to determine the outcome. The last season 2018-2019 is used as a test season to evaluate the strength of the approach.

Even though the NBA and EuroLeague play the same game with the same rules, both leagues are quite different. They have a different format, different number of teams and changes in teams competing, different number of games, different styles of plays, and different seasonal trends. In general, the NBA is a more fast-paced competition, with proportionally more plays in total per game, and more zero-, one-, and two-point plays than the EuroLeague. The NBA is gradually becoming even more fast-paced due to the drastically increasing popularity of three-pointers. The EuroLeague has always proportionally had more three-point plays than the NBA, until it was finally overtaken by the NBA in the last two seasons.

The models predicting the different types of plays were not very accurate but did improve the prediction accuracy of the simulations by 3% over simply taking the average and provided insight in the factors influencing the different types of plays. Home advantage plays an important role in both leagues, especially in the number of total plays, zero-point plays and two-point plays, but seemed to be more important in the EuroLeague than in the NBA. Generally, the team's opponent is most important in predicting the different types of plays.

Playing against fast-paced opponents will result in more plays for the team overall, while the opposite is true for playing against an opponent who gets a lot of offensive rebounds. Playing against defensive teams will result in more zero-point plays and less two- and three-point plays. Playing against aggressive teams committing a lot of personal fouls will result in more one-point plays.

Finally, the simulations performed very well in predicting the game's winner in both leagues. In the NBA, 65.4% of all games over all seasons and 67.3% of the games in the test season were predicted correctly. In the EuroLeague, these numbers were respectively 71.1% and 78.6%. These prediction accuracies are similar, if not better than the prediction accuracies of other papers using various machine learning algorithms and only slightly less accurate than the betting market and the experts. The strength of the approach over other approaches is however its simplicity, interpretability, ease of use, and ease of visualisation.

2 Introduction

The most common, and perhaps the most important question, asked in any sport at any level is: Who will win? That question is followed by an arguably even more important question: Why will that team win? Those two questions can of course rarely be answered with 100% certainty, otherwise there wouldn't be any of the excitement and anticipation we feel around sports. Yet, they are nonetheless two questions everyone will always try to find the best answer to. This answer is sought to get an edge over your opponents, be it your opponent in the game, the people betting against you, or just your discussion partner in a bar. In this paper, we will take a shot at answering these questions for basketball games played in the NBA and the EuroLeague in the last 19 seasons.

The rise in importance of statistics and mathematical modelling in sports within the last decades means a lot of effort has already been put into statistically predicting the outcome of basketball games. Various modelling techniques have been proposed to predict the winning team. Some examples of methods used are k-nearest neighbours [1], naive Bayes classifier [2][3][5], probit regression [4], decision trees and random forests [5], support vector machines [6], neural networks [5][6] [7], and linear and logistic regression [7]. These models are usually based on the traditional box score summary statistics that are commonly recorded for each game, and aim to classify each opposing team as winner or loser.

An important change in the last decade has been the introduction and public availability of play-by-play data. Play-by-play data contains a much more detailed description of the events happening in a basketball game compared to traditional box score statistics. Multiple studies have already exploited this additional information to forecast the outcome of basketball games by simulating them using a Markov chain model [8][9][10], an alternating renewal-reward process model [11], continuous-time anti-persistent random walk model [12], or a Poisson regression [13]. In this paper we will also try to simulate basketball games to predict their outcome.

Central to these models is the concept of possessions. Possessions, although already known and used before, were popularized in basketball analytics by Dean Oliver in his book "Basketball on Paper" [14], and currently form the basis of a lot of advanced basketball analytics work. The concept is simple and attractive. Each possession of a team ends when the other team gains possession of the ball.

Therefore, two opposing teams have the same number of possessions in the game, and the way they utilize their possessions determines the better team. It is not an officially tracked statistic and in many cases is still estimated from the box score statistics using the formula proposed by Dean Oliver [14], yet it can also be directly counted from play-by-play data.

A similar concept, yet barely used in basketball analytics compared to possessions, is the concept of plays. A play is essentially the same as a possession, except that a new play starts after an offensive rebound, while this does not initiate a new possession. This does not have the same attractive property of being approximately the same for each opposing team and is sometimes deemed not useful as a basis for evaluating team efficiency [15]. It is, however, more interesting for our purpose as it contains more information than the concept of possessions, which is theoretically flawed. A team can score an unlimited amount of points in a single possession if it keeps on getting fouled after a scored field goal, misses the free-throw and gets the offensive rebound. Additionally, the team can play the whole game in a single possession if it also gets the offensive rebound after each failed shot. This never happens of course, which makes analysis based on possessions so relevant, but interesting sequences have been recorded. The Miami Heat recorded a 90 second possession against the Brooklyn Nets in the Eastern Conference Finals in 2014, with four missed field goals and four offensive rebounds, eventually scoring a lay-up [a]. Larry Bird's Boston Celtics had a 72 second long possession in a 99-99 game 7 tie against the Detroit Pistons in the 1987 finals, with four minutes left in the 4th quarter, getting 5 offensive rebounds and eventually scoring a three-pointer [b]. These sequences are much more than a single two- or three-point possession and are much better described as four or five zero-point plays and a two- or three-point play. Another example of an aberration is the eight-point possession of the Golden State Warriors against the Portland Trailblazers on the 13th of February 2019 [c].

Following the concept of plays, a team can score zero, one, two, three, or four points per play. If we can predict the number of plays each opposing team will make in a game, and predict which percentage of these plays will result in zero, one, two, three, or four points, we can not only predict the winner of the game but also how many points each team scored. Additionally, as we are predicting the different kind of plays for both EuroLeague and NBA, we can identify their most influential factors and compare them between both leagues. Since the frequency of scoring can be found to obey a Poisson-like process [12][13], we use a Poisson regression to predict the different number of plays. We then use a compound Poisson process to simulate each game to find the answers we are looking for.

3 Methods

3.1 Summary

For both the NBA and EuroLeague, freely available box scores, schedules, and play-by-play data from 19 seasons were scraped using custom R scripts and available R packages. The raw play-by-play data was processed using regular expressions into a table with the number of total plays, zero-point plays, one-point plays, two-point plays, three-point plays, and four-point plays, as well as the number of possessions, for each opposing team, for each game. The rest of the scraped data was then wrangled into a comparable format containing the same information for each league. Validity of the data, outliers, and missing data were checked at multiple points during the process.

Seasonal trends in the different play data as well as correlations between the plays data and other variables were explored within each league separately before running prediction models. After exploring the data, penalized Poisson regression models using Elastic Net regularisation were performed to predict the total number of plays by each team, per game, as well as the other five types of plays. For each league and each type of play, one general model was first fitted on the data comprising all teams and all seasons but the last. Separate models were then fitted for each team for the first 18 seasons. The significant variables could then also be compared between both leagues and between teams. The models were then applied to games played in the last season, 2018-2019, to predict the different kinds of plays. Lastly, the outcome of each game, using the results from the models, was simulated using a Compound Poisson model to predict the outcome of each in terms of final score.

3.2 Data Collection

All the data used for this research was scraped from two websites:

- https://www.basketball-reference.com/
- https://www.euroleague.net/

For both leagues, all the data available for all games from the 2000-2001 season up to and including the 2018-2019 season was scraped. From the gathered data, three types of data were eventually used in the analysis:

a. Schedule and Results
    - Date
    - Home Team
    - Home Points
    - Visiting Team
    - Visiting Points
    - Overtimes
    - Regular Season/Playoffs

    This data came from the basketball reference site for both leagues. In total 24531 games were recorded in the NBA and 4334 games in the EuroLeague in those 19 seasons.

b. Box Scores
    - Home Basic Box Scores
    - Visitors Basic Box Scores
    - Home Advanced Box Scores (Only for NBA)
    - Visitors Advanced Box Scores (Only for NBA)

    The box scores for the NBA came from the basketball reference site. The site, however, has only box scores data for the EuroLeague from the 2014-2015 season on. Therefore, this data was gathered from the official EuroLeague site. For two games in the EuroLeague the box scores were missing, one in the 2000-2001 season and one in the 2001-2002 season. These games were excluded from the analysis.

c. Play-By-Play

    This data is a recollection of all the events that happened during the game, written out as single sentences. These events include:
    - Jump balls
    - Timeouts
    - Defensive/Offensive rebounds
    - Fouls
    - Missed/Made free throws
    - Missed/Made two-point shots
    - Missed/Made lay-ups/dunks

- Missed/Made three-point shots
- Turnovers
- Player changes

This data was gathered from the basketball reference site for the NBA and from the official EuroLeague site for the EuroLeague. For the EuroLeague, play-by-play data was available from the 2006-2007 season on, with the exception of one game in the 2003-2004 season. Additionally, play-by play data was missing for 1 game in the 2007-2008 season, 43 games in the 2008-2009 season, two games in the 2010-2011 season, five games in the 2013-2014 season, one game in the 2014-2015 season and three games in the 2016-2017 season.

All the data was scraped using custom R scripts[1] using the Rvest package, except for the EuroLeague for which play-by-play data was obtained using the extractPbp() function of the Eurolig package by Sergio Olmos Pardo[16].

3.3 Data Wrangling

After extracting the data, all the data was wrangled into a useable format and combined into one data frame. First, the three types of data are processed separately.

3.3.1 Schedule and Results

A game ID was created for each game. For the NBA games, this game ID was the game ID used by the site to reference to the game. It was extracted from the html code of the basketball reference site together with the table. The format of the game ID is "YYYYMMDD0XXX", with "YYYY" the year, "MM" the month, and "DD" the day the game was played, followed by a zero, and then the three-letter abbreviation of the home team in upper case (e.g. 201801020CLE). For the EuroLeague, a similar format "YYYY-MM-DD_xxxx" was employed. Here the abbreviation is in lower case and has a variable length. Most of the abbreviations come from the basketball reference site. The abbreviations used for each team can be found in Appendix 1.

1 First script written was to extract NBA schedule and results from basketball-reference.com, and was an adaptation of a script found on Github: https://github.com/kjytay/misc/blob/master/blog/2018-12-11_nba_game_data.R. The other scripts are entirely custom scripts.

Special care was taken in creating the abbreviations because of several name changes of the franchises over the years. This is especially relevant for the EuroLeague, as the teams playing change every year, can come from the same city, and frequently have sponsor names in their teams name. For example, Climamio Fortitudo Bologna and Skipper Bologna are the same team, but Kinder Bologna is a different team.

3.3.2 Box Scores

Player specific statistics where dropped from the box scores and only team statistics per game were kept. Statistics were added, adapted or removed to have the same statistics for both the NBA and the EuroLeague games, as both leagues record game statistics differently. For example, the NBA records the number of field goals attempted or scored, which is the sum of the two-point shots (lay-ups and dunks included) and three-point shots but does not record two-point shots separately. The EuroLeague records the number of two-point shots and three-point shots but does not sum them together in a field goals statistic. The EuroLeague also does not specify any advanced box scores. These are summary statistics derived from the basic box scores. They were calculated and added for the EuroLeague games[2]. A game ID was again created for each game, corresponding with the game ID of the game in the schedule data.

For the full list of statistics, see Appendix 2.

3.3.3 Play-By-Play

Before going any further, a formal definition of a possession and a play needs to be established.

- "A *possession* starts when one team gains control (or possession) of the basketball and ends when that team gives up control of the basketball. Teams can give up possession of the basketball in several ways, including (1) made field goals or free throws that lead to the other team taking the ball out of bounds, (2) defensive rebounds, and (3) turnovers." [17]

2 Calculated according to the formulas on https://www.basketball-reference.com/about/glossary.html

- A *play* starts when one team gains control of the basketball after 1) made field goals or free throws that lead to the other team taking the ball out of bounds, (2) defensive rebounds, (3) turnovers, and (4) offensive rebounds.

The difference lies thus in that an offensive rebound starts a new play but does not start a new possession. Therefore, two teams in a game will always have approximately the same number of possessions, while the number of plays can differ greatly. This concept of approximately equal possessions for two teams in a game has played a central role in basketball analytics, notably in evaluating the efficiency of teams and individuals [17][18]. The term (minor) possession can also be used for referring to a play. To avoid confusion, possessions will be referred to as team possessions and plays as plays in this paper[3].

For both teams in each game, the following information was extracted from the raw play-by-play data:

1. *Zero-point plays (0PP)*: Number of plays resulting in zero points for the team. A zero-point play can result from 1) a turnover, 2) a missed field goal, 3) missed free throw after a technical foul, and 4) missing all the free throws after a foul resulting in two or three free throws.
2. *One-point plays (1PP)*: Number of plays resulting in one point for the team. A one-point play can result from 1) a scored free throw after a technical foul, or 2) a single scored free throw after a foul resulting in two or three free throws.
3. *Two-point plays (2PP)*: Number of plays resulting in two points for the team. A two-point play can result from 1) a scored two-point field goal, 2) a missed free throw after a fouled and made two-point field goal, and 3) scoring two free throws after a foul resulting in two or three free throws.
4. *Three-point plays (3PP)*: Number of plays resulting in three points for the team. A three-point play can result from 1) a three-point shot made, 2) a scored free throw after a fouled and made two-point field goal, 3) a missed free throw after a fouled and made three-point shot, and 4) scoring three free throws after a fouled three-point shot.
5. *Four-point plays(4PP)*: Number of plays resulting in four points for the team. A four-point play can result from 1) a scored free throw after a fouled and made three-point shot.

3 During my analysis I still referred to plays as possessions. When a possession is mentioned in the variable names, or the comments in the R scripts, it refers to a play, unless it specifically mentions a team possession.

6. *Total plays (TP)*: Total number of plays made by the team in the game. This is the sum of the five kinds of plays described above.
7. *Team possessions (TMP)*: The number of times a team gains possession of the ball and corresponds to a possession as described above.

After processing, the results were compared with the box score data of the game to check for discrepancies. The weighted sum of the plays was compared to the points scored by the team according to the box scores:

$$PTS = \sum(0 * 0PP + 1 * 1PP + 2 * 2PP + 3 * 3PP + 4 * 4PP)$$

This serves as a double check to make sure that both the plays and the box scores of the game were recorded correctly. Discrepancies in the processed data were found for 44 of the 24531 NBA games and for 16 of the 2946 EuroLeague games (Appendix 3). From experience collecting and processing the play-by-play data, and debugging the script scraping the data, most of these discrepancies are caused by inconsistent notation in the play-by-play records. Typically, these involve (technical) free throws, as reflected by the differences being mostly only one point. The data is adjusted to correct these discrepancies by adding/removing a 1PP, a 2PP, and a 1PP and 2PP for the one-point, two-points, three-points differences respectively. The impact of manually changing this data should be minimal, as those games are randomly spread between season and teams. However, it should be kept in mind in terms of bias because we cannot be sure that these are the correct possessions to attribute the differences to. Discrepancies of more than three points were examined and adjusted manually.

3.3.4 Full Data

After processing the three types of data, they were combined. For each league, an overall data frame was created containing all the games played during all the recorded seasons. Additionally, a separate data frame was created for each team, containing only the games played by that team. As mentioned above, special care was taken in identifying teams that changed their name, to not falsely identify new teams.

Each game was split into two observations, one for each opposing team. Statistics corresponding to the home team and the visiting team were renamed to team and opponent statistics. The team statistics for one observation correspond to the statistics of the home team, while the opponent statistics correspond to the statistics of the visiting team.

The team statistics of the other observation then correspond to the statistics of the visiting team, and the opponent statistics to the home team. Both observations contain exactly the same information. The data is effectively duplicated, which makes it easier in subsequent analysis to split the data by team and focus the analysis on team/opponent statistics.

Multiple discrepancies were found between the schedule and the box scores data for the EuroLeague. When resolving these discrepancies, the data found on the official EuroLeague site was assumed correct. In total, 110 discrepancies were found. Most were dates that were incorrect in the schedule data, possibly due to the difference in date notation between Europe (EuroLeague site) and the United States of America (basketball reference site). Others involved differences in game outcomes, switching home- and away team, and games with no records. For five games, the final score and the box scores statistics on the official EuroLeague site did not coincide. The final score was assumed correct in those cases because a human mistake in attributing a point to a player on the score sheet is easier done than not attributing points to a team after scoring with the whole stadium watching. In two of those five games, play-by-play data was available to confirm that the final score was correct and the box scores incorrect.

Finally, the play-by-play data was explored to evaluate outliers, observe missing values, and identify seasonal trends using five-number summaries and box scores.

3.3.5 Average Data

We now have the raw data for each game. Most of this data is, however, not yet available when the purpose is predicting the number of different plays of each opposing team during a game. The only data known about a game before it is played, is the type of game (Regular Season or Playoffs), date and time of the game, the teams playing, home advantage, and statistics of previous games of each team. To include statistics from previous games played by both teams, a summary statistic has to be employed for that data. We need to take into account the differences between both leagues.

The franchises competing in the NBA are always the same, except between 2000 and 2004 when one franchise was missing (Table 1). Also, the number of games played was always the same, except in the season 2011-2012.

Both the number of teams competing in the EuroLeague, as the number of regular games played per team, have changed multiple times during the last 19 seasons (Table 2). The total number of games played during one season is also much lower than in the NBA.

Season	Total Games	Teams	Regular Games
2000-2001	1189	29	82
2001-2002	1189	29	82
2002-2003	1189	29	82
2003-2004	1189	29	82
2004-2005	1230	30	82
2005-2006	1230	30	82
2006-2007	1230	30	82
2007-2008	1230	30	82
2008-2009	1230	30	82
2009-2010	1230	30	82
2010-2011	1230	30	82
2011-2012	990	30	66
2012-2013	1229	30	82
2013-2014	1230	30	82
2014-2015	1230	30	82
2015-2016	1230	30	82
2016-2017	1230	30	82
2017-2018	1230	30	82
2018-2019	1230	30	82

Table 1. Number of teams and number of regular games played in each NBA season.

Season	Total Games	Teams	Regular Games
2000-2001	120	24	10
2001-2002	224	32	14
2002-2003	168	24	14
2003-2004	168	24	14
2004-2005	168	24	14
2005-2006	168	24	14
2006-2007	168	24	14
2007-2008	168	24	14
2008-2009	120	24	10
2009-2010	120	24	10
2010-2011	120	24	10
2011-2012	120	24	10
2012-2013	120	24	10
2013-2014	120	24	10
2014-2015	120	24	10
2015-2016	120	24	10
2016-2017	240	16	30
2017-2018	240	16	30
2018-2019	240	16	30

Table 2. Number of teams and number of regular games played in each EuroLeague season.

Next to the changes mentioned above, players change teams regularly between two seasons, and even during a season. The performance of a team is thus generally not comparable between seasons. To take all this into account, the summary statistic used for a team's box score and play-by-play statistics is the mean of the previous games played by the team in the same season. This means that for the first game of the season of each team no statistics are available, for the second game the statistics are the statistics of the first game, the statistics of the third game the average of the first two games, and so forth.

Lastly, a few more variables were added relating to the number of wins and losses of each team, and also the number of consecutive road games a team is playing. For a full list of the used variables, see Appendix 4.

3.4 Analysis

A penalized Poisson regression model using Elastic Net regularization was constructed to predict the number of zero-point plays, one-point plays, two-point plays, three-point plays, four-point plays, and total plays in each league.

The model was first fitted on the pooled data from all teams and the first 18 seasons to evaluate general predictive power and compare the overall trends between both leagues. Secondly, separate models were fitted for each team, including a model with an interaction term between season and opponent, to compare the fitted models between teams and possibly increase prediction accuracy. These models were also fitted using the data of the first 18 seasons. The models were then applied to predict the different kinds of plays of all the games in the last season. Finally, a Compound Poisson process was utilized to simulate each game 10'000 times, using the predicted number of plays, in order to predict the outcome of each game.

3.4.1 Poisson Regression

When trying to predict the number of plays (0PP, 1PP, 2PP, 3PP, 4PP, and TP), we are analysing count data. Count data gives the number of occurrences of a certain event over a specific period of time, which in this case is the number of plays during the length of a basketball game. Using ordinary least squares (OLS) regression on count data, especially when the mean of the outcome is relatively low (lower than ten as a rule of thumb), can cause undesirable results, including biased standard errors and significance tests [19]. This is the case for the number of 1PP, 3PP and 4PP, which regularly have a mean outcome of less than ten. This is due to the inherent heteroscedasticity of count data, and the fact that its conditional distribution tends to be positively skewed and have positive kurtosis, with many low count observations and none below zero [19]. A better alternative is a Poisson regression. Poisson regression is a member of the generalized linear models (GLM) family with Poisson distribution error structure and the natural log (ln) link function [19][20][21]. Given an outcome variable Y with a Poisson distribution whose mean depends on p predictor variables X

$$E(Y|X) = \hat{\mu}$$

the Poisson regression model can be written as

$$\ln(\hat{\mu}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

where $\hat{\mu}$ is the predicted count of the outcome variable given the specific values of the predictors $X_1, X_2, \ldots, X_p$, and the βs the regression coefficients, with β_0 the intercept. The log-likelihood function for n observations $\{x_i, y_i\}_1^N$ is then given by [20][22]

$$l(\beta|X,Y) = \sum_{i=1}^{n} \left(y_i(\beta_0 + \beta' x_i) - e^{\beta_0 + \beta' x_i}\right)$$

where x_i is the vector of all the predictor variables $(X_1, X_2, \dots, X_p)$ for observation i, y_i is the actual outcome of the observation i, n is the number of observations, β_0 a given intercept, and β' a given vector of parameters. This log-likelihood function is used to derive the maximum likelihood estimator of the vector of parameters β $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$. This estimator $\hat{\beta}$ is obtained by solving

$$\hat{\beta} = \arg\max_{\beta \in R} l(\beta|X,Y)$$

or in other words, by finding the parameters $\hat{\beta}$ that maximizes the log-likelihood of the observations X, Y.

3.4.2 Elastic Net Regularisation

The number of variables considered in the prediction model is high, especially for the models where an interaction term between season and opponent is added. This can cause computational and statistical problems for the more basic variable selection methods such as forward or backward stepwise selection. In multiple models for the EuroLeague, the number of predictor variables exceeds the number of observations, $p > n$. This means that there is no unique least squares estimate anymore and the variance of the coefficients increases dramatically. Additionally, multiple predictor variables are strongly correlated with each other causing multicollinearity problems. A solution to these problems is using regularisation by adding a L1- or/and a L2 penalty term to the log-likelihood function.

The *L2-penalty term (ridge penalty)* shrinks the coefficients of correlated variables towards each other and asymptotically towards zero, allowing them to borrow strength from each other. It can however not perform variable selection as it always keeps all the variables on the model (coefficients do not become zero) [23][24][25].

The *L1-penalty term (lasso penalty)* performs both variable selection and regularisation. It will select a subset of the provided variables to fit the final model by shrinking some coefficients to zero. Lasso handles correlated variables differently than ridge by selecting one of the correlated variables and shrinking the others to zero, without caring much about which variable is selected.

In other words, ridge will use information from all the individually correlated variables, and shrinks their individual absolute importance, while lasso will take one of the correlated variables at random and only uses the information contained by that variable, without shrinking its importance. Another limiting feature of the lasso is that in cases where $p > n$ it selects at most n variables before it saturates [23][24][25]. Being able to select more than n variables could be necessary when creating models for EuroLeague teams that played a very small amount of games (e.g. a team that only played in one EuroLeague season with ten regular season games).

A compromise between these two penalties is the *elastic-net penalty*, which uses both the L2- and L1- penalty terms and is particularly useful when $p > n$ and when many variables are correlated. The elastic-net penalty is given by [23][24]

$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}||\beta||_{l_2}^2 + \alpha\,||\beta||_{l_1}$$

$$= \sum_{j=i}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right]$$

where $||\beta||_{l_2}^2$ is the L2-penalty term and $||\beta||_{l_1}$ the L1-penalty term. When $\alpha = 1$ the elastic net penalty becomes the ridge penalty and when $\alpha = 0$ the elastic net becomes the lasso penalty.

3.4.3 Penalized Poisson Regression

The penalized Poisson regression is then defined by [22][25]

$$PPR = l(\beta|X,Y) + \lambda\, P_\alpha(\beta)$$

$$= \sum_{i=1}^{n}\left(y_i(\beta_0 + \beta' x_i) - e^{\beta_0+\beta^T x_i}\right) + \lambda\sum_{j=i}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right]$$

where $\lambda \geq 0$ is the tuning parameter, which controls the strength of the shrinkage, or weight, of the penalty term. The value of λ depends on the data and it can thus be calculated via cross-validation [25]. Optimization of this penalized log-likelihood is then achieved by

$$min_{\beta_0,\beta}\ -\left[\frac{1}{n}\,l(\beta|X,Y)\right] + \lambda\left((1-\alpha)\sum_{j=1}^{p}\frac{1}{2}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right)$$

which is the formula used in the glmnet R package used for the analysis [22]. They standardise the log-likelihood term during the optimization by dividing by the number of observations n. Their algorithm uses a quadratic approximation to the log-likelihood, and then coordinate descent on the resulting penalized weighted least-squares problem. These constitute an outer and inner loop. They use an outer Newton loop and an inner weighted least-squares loop to optimize this criterion [22].

3.5 Evaluation measures

3.5.1 MSE

The two most commonly used performance measures for the Poisson regression are the deviance (D) and the mean-squared error (MSE). The MSE is the mean squared difference between the estimated value and the real value, given by the formula

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

where n is the number of observations, y_i is the observed value of observation i, and $\hat{\mu}_i$ the estimated value for observation i.

A smaller MSE equals a better fit of the model. A variant of this measure is the root-mean-squared error (RMSE), which is just the square root of the MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}$$

This measure gives an idea of the average deviation of the predicted value $\hat{\mu}_i$ from its observed value y_i. The effect of each error on the MSE, and consequentially on the RMSE, is proportional to its size, meaning larger errors will have a larger effect on the RMSE than smaller errors, making the RMSE sensitive to outliers. As mentioned above, count data has inherent heteroscedasticity and a positively skewed conditional distribution, especially in count data with a low mean, which can inflate the RMSE. It is however still a frequently used cost function and generally performs well when the distribution is close to a normal distribution.

3.5.2 Deviance

A strictly proper cost function for count data is the deviance (D). The deviance is a measure of goodness of fit for generalized linear models (GLM), and a generalization of the idea of using the residual sum of squares in ordinary least square (OLS) regression to cases where model-fitting is done by maximum likelihood. It is defined as minus two times the difference between the maximized log-likelihood (MLE) of the (unsaturated) model and the maximized log-likelihood of the saturated model [20][21][26]. "A saturated model has a separate parameter for each observation y_i. It gives a perfect fit." [26].

$$D(y;\,\hat{\mu}) = -2\,log\left(\frac{maximum\ likelihood\ model}{maximum\ likelihood\ saturated\ model}\right)$$

$$= -2\,log\left(\frac{L(\hat{\mu};\,y)}{L(x;y)}\right)$$

$$= -2[\log\big(L(\hat{\mu};\,y)\big) - \log\,(L(x;y))]$$

$$= -2\,[l(\hat{\mu};y) - l(x;y)]$$

$$= 2\,[l(x;y) - l(\hat{\mu};y)]$$

Here, y is the vector of observed values, $\hat{\mu}$ the vector of parameter values for μ giving the maximum (log) likelihood of the model, and x the vector of parameters values for the saturated model, with a separate parameter for each observation, giving a perfect fit. The log-likelihood function l is the logarithm of the likelihood function L.

The random component of the GLM specifies that the N observations $(y_1, \ldots, y_N)$ on Y are independent, with probability mass or density function for y_i of the form:

$$\mathrm{f}(y_i;\theta_i,\phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)\}$$

This is called the exponential dispersion family and ϕ is called the dispersion parameter. The parameter θ_i is the natural parameter.

When Y_i is Poisson,

$$f(y_i;\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp(y_i\log(\mu_i) - \mu_i - \log y_i!)$$

$$= \exp[y_i\theta_i - exp(\theta_i) - \log y_i!]$$

where $\theta_i = \log(\mu_i)$. This has an exponential dispersion form with $b(\theta_i) = exp(\theta_i)$, $a(\phi) = 1$, and $c(y_i, \phi) = -\log y_i!$. The natural parameter is $\theta_i = \log(\mu_i)$, therefore $b(\theta_i) = exp(\theta_i) = exp(\log(\mu_i)) = \mu_i$.

Going back to the deviance, we get

$$D(y;\ \hat{\mu}) = 2\ [l(x;y) - l(\hat{\mu};y)]$$

$$= 2\sum_{i}^{n}\left[y_i\tilde{\theta}_i - b(\tilde{\theta}_i)\right]/a(\phi) - \ 2\sum_{i}^{n}\left[y_i\hat{\theta}_i - b(\hat{\theta}_i)\right]/a(\phi)$$

$$= 2\sum_{i}^{n}\left[y_i(\tilde{\theta}_i - \ \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right]/a(\phi)$$

As seen before, for Poisson GLMs, $\hat{\theta}_i = \log(\hat{\mu}_i)$, and $b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$, for the unsaturated model. Similarly, for the saturated model $\tilde{\theta}_i = \log(y_i)$ and $b(\tilde{\theta}_i) = \exp(\tilde{\theta}_i) = \ y_i$. Also $a(\phi) = 1$, so the deviance equals to

$$D(y;\ \hat{\mu}) = \ 2\sum_{i=1}^{n}[y_i\log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$$

When a model with log link contains an intercept term, the likelihood equation implied by that parameter is $\sum y_i = \sum \hat{\mu}_i$, which simplifies the deviance to

$$D(y;\ \hat{\mu}) = \ 2\sum_{i=1}^{n}[y_i\log(y_i/\hat{\mu}_i)]$$

The deviance function used in the glmnet R package for the analysis seems to be a normalised version of the function above:

$$\frac{1}{n}D(y;\ \hat{\mu}) = \ \frac{1}{n}2\sum_{i=1}^{n}[y_i\log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$$

Using this formula gives approximately the same deviance as returned by the glmnet R package. Because log (0) is undefined, when an observed value y_i is equal to zero, $y_i\log(y_i/\hat{\mu}_i)$ is set to zero.

3.5.3 R-squared

A third performance measure we used is the R^2, or at least a pseudo-R^2, for the Poisson regression. For the ordinary least squares (OLS) the R^2 is a measure for the proportion of variation in the outcome that is accounted for by the predictors and is measured in terms of sum of squares [19]. In regression with multiple predictor variables the R^2 is adjusted for the number of predictor variables k included in the model

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-k}}{\frac{SST}{n-1}}$$

where SSE is the error sum of squares, n the number of observations, k the number of parameters in the model, and SST the total sum of squares.

The difference between the deviance or RMSE, and the R^2 is that the R^2 can be interpreted as a standalone value. It gives a value between zero and one indicating the percentage of additional variation is explained by the model compared to a model with intercept only. The former two measures are relative measures that can only be interpreted in relation to another model that uses the same parameters and comparable data (e.g. a model with intercept only which returns the R^2 value, or a model with different parameters to choose the better model). The main issue is that, for the Poisson regression the total variation in the outcome cannot completely be divided into explained and unexplained parts. The pseudo-R^2 used is a measure that gives the proportional reduction in deviance by including predictor variables compared to a model with intercept only [19][21].

The deviance of the model to test is, as seen before

$$D(y;\ \hat{\mu}) = \ 2\sum_{i=1}^{n}[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$$

where y_i is the observed value and $\hat{\mu}_i$ the predicted value for observation i. The deviance of a model with only an intercept is

$$D(y, \overline{y}) = \ 2\sum_{i=1}^{n} y_i \log(y_i/\overline{y})$$

where $\overline{y}$ is the mean of all observed values y.

This gives the deviance R^2 for Poisson

$$R^2_{DEV} = 1 - \frac{2\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}}{2\sum_{i=1}^{n} y_i \log(y_i/\bar{y})}$$

which satisfies all the conditions required for a good R^2measure. This formula is basically the equivalent of the R^2_{adj} of the OLS regression. The numerator is the deviance of the model, which in the case of OLS is captured by $MSE = \frac{SSE}{n-k}$, while the denominator is the deviance of the model with only the intercept, $\frac{SST}{n-1}$ for OLS regression. Note that the log of zero is not defined and you cannot divide by zero. The observed values y_i that are equal to zero are replaced by 10^{-10}.

3.6 The Models

3.6.1 General Model

First, a model was created to predict each number of plays using data from the first 18 seasons for each respective league. The data was first split into a training and test set in an 80/20 ratio. All the numeric predictor variables were scaled in each set using their mean and standard deviation in the training set, and dummy variables were created for the non-numeric predictor variables where necessary. A tuning grid was created to train for the best value for α in the elastic net penalty. Values ranging from zero to one in steps of 0.1 were evaluated. For each value of α, the tuning parameter λ was trained using five-fold cross-validation on the training set. The model with the optimal value for λ, the model with the lowest deviance, was then selected to be the optimal model for that value of α. The value for α which optimal model had the lowest deviance was then finally chosen. The model was then evaluated both on the training and the test set using the three evaluation measures (RMSE, D, R^2_{DEV}), and the significant coefficients were plotted in order of importance. The final model was then used to predict the different kinds of plays for all the games in the last season (2018-2019).

3.6.2 Team Models No Interactions

The same protocol was applied to create a model for each separate team (Team Models No Interactions 1).

Tuning the α of the elastic net penalty generally returned an optimal value of $\alpha = 0$ or $\alpha = 0.1$. A value of $\alpha = 0$ means the elastic net penalty is effectively the ridge penalty and no variable selection is performed.

Variable selection is important, therefore the models for each team were trained again but this time without tuning the α parameter but setting it to 0.1 (Team Models No Interactions 2). This also saved considerable amounts of computing time. Multiple teams did not have any 4PP in their EuroLeague history, or too few to reach convergence for cross-validation. No model was fitted for predicting the 4PP of those teams and all predicted values were set to zero for the Compound Poisson simulation. Some teams playing in the 2018-2019 EuroLeague season did not play any EuroLeague games in one of the previous seasons, hence no team model could be constructed to predict their plays. The average of each play over their previous games played in the season was taken to replace the model prediction for these teams.

3.6.3 Team Models Interactions

A third set of models was again fitted for each team using the same protocol, with $\alpha = 0.1$, but this time with an interaction term added between the variables season and opponent.

3.6.4 No Prediction Models

To evaluate the value of predicting the different types of plays to use in the simulations, a last dataset was set up. In this dataset, the different types of plays used in the simulations are not predicted using complex predictions models but are simply the average number of TP, 0PP, 1PP, 2PP, 3PP, and 4PP of the team so far that season.

3.7 Simulations

Finally, each game was simulated 10’000 times using a Compound Poisson process,

$$Y = \sum_{i=1}^{T} X_i$$

where T is the number of events following a discrete distribution, and X_i the magnitude of the outcome of the i_{th} event [27]. Y is the points scored by a team during the game.

T is the total number of plays during the game by that team which follows a Poisson distribution with mean equal to the predicted total plays (TP). X_i is the number of points scored per play (0, 1, 2, 3 or 4), with the outcome of each play randomly sampled from a multinomial distribution with k = 5, one for each possible type of play (0PP, 1PP, 2PP, 3PP, 4PP), and their probabilities the ratios of their predicted amount over the predicted total number of plays.

$$f(x_1, x_2, x_3, x_4, x_5; n, p_1, p_2, p_3, p_4, p_5)$$

$$= f\left(0PP, 1PP, 2PP, 3PP, 4PP; TP, \frac{0PP}{TP}, \frac{1PP}{TP}, \frac{2PP}{TP}, \frac{3PP}{TP}, \frac{4PP}{TP}\right)$$

$$= \begin{cases} \frac{TP!}{0PP!\,1PP!\,2PP!\,3PP!\,4PP!} \left(\frac{0PP}{TP}\right)^{0PP} * \left(\frac{1PP}{TP}\right)^{1PP} * \left(\frac{2PP}{TP}\right)^{2PP} * \left(\frac{3PP}{TP}\right)^{3PP} * \left(\frac{4PP}{TP}\right)^{4PP} & when \sum_{i=1}^{k} x_i = n \\ 0 & otherwise \end{cases}$$

Figure 1 shows an example of how a single simulation looks like for one NBA team. In this simulation, the team has 95 possessions over the course of a 48 minutes game and scores 81 points. The sequence of plays follows a Poisson process, as reflected by the distribution of the points along the x-axis. The outcome of each play is sampled from a multinomial distribution, and their value shown by the colour of each individual point. The score of the team at a certain time in the game (x-axis) is represented by the cumulative value of the points and is shown on the y-axis.

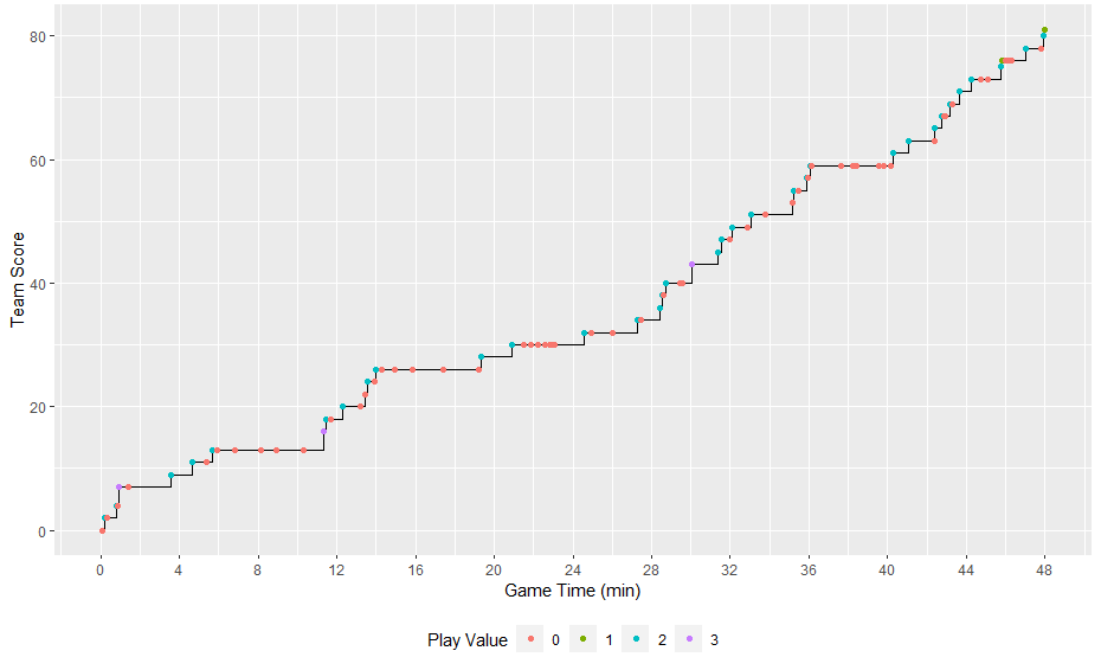


Figure 1. Example of a single simulation for a NBA team. The team scored 81 points over 48 minutes in 95 possessions. The colour of each point represents the outcome of the play (orange = 0PP, green = 1PP, blue = 2PP, purple = 3PP).

The end result of each of these 10’000 simulations, the amount of points scored by the team at the end of the game, is then plotted as a density function to determine the most likely outcome. The same is done for the opponent. The predicted winner of the game is the team that won the most simulations.

This process was done five times for each league, one time using the predicted data coming from the model that uses all the data available, one time for each of the three groups of team models, and one time using the averages. To evaluate which model worked best, the prediction accuracy was measured. First, for each simulation of a single game, the predicted winner was determined. The percentage of games this predicted winner was the actual winner is then the prediction accuracy of the simulation. The accuracy was measured for the training seasons (2000-2018) and the test season (2018-2019) separately and together.

4 Results

4.1 Data Exploration

4.1.1 Teams

In total, 24531 NBA games were recorded from season 2000-2001 up to and including season 2018-2019. Each franchise played 1558 regular games in the last 19 seasons, except the now New Orleans Pelicans (NOH) and the now Charlotte Hornets (CHA) (Fig. 2). On top of the regular games, each franchise played a certain amount of playoff games, with the San Antonio Spurs (SAS) clearly being the most successful franchise in terms of playoff games played, and the New York Knicks (NYK) the least successful.

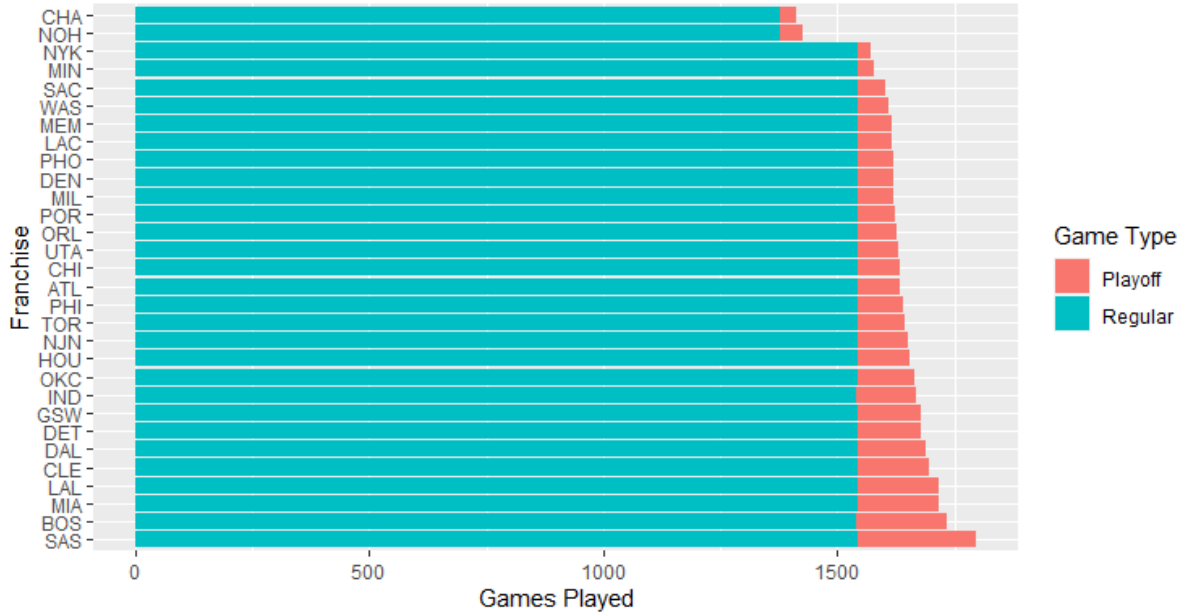


Figure 2. Games played per franchise in the NBA from season 2000-2001 up to and including season 2018-2019. The blue bars show the number of regular season games and the red bars the number of playoff games played.

Unlike in the NBA, in the EuroLeague different teams play against each other each year. Figure 3 shows that there were 84 different teams playing in the EuroLeague in the 19 seasons recorded. Only 61 of those teams have any recorded play-by-play data. Žalgiris, Saski Baskonia (vitoria), FC Barcelona Bàsquet, and Olympiacos B.C. are the teams with the most regular games in the EuroLeague. The team with the most EuroLeague games and the most playoff games is PBC CSKA Moscow. Other important teams that played comparable amounts of regular and playoff games are Panathinaikos B.C., Maccabi Tel Aviv B.C., Real Madrid Baloncesto, Anadolu Efes S.K., Fenerbahçe Basketball, followed by Unicaja Baloncesto Malaga, and Olimpia Milano. These teams, except Olimpia Milano, played between 300 and 470 EuroLeague games in total in the last 19 seasons.

On the other end of the spectrum are teams that only competed in the EuroLeague during a single season, without going to the playoffs. These teams played only 10 to 14 games in total, depending on the season they were competing in. Examples of these teams are KK Zagreb, PGE Turów Zgorzelec, BC Zenit Saint Petersburg, Chorale Roanne Basket, and more (Fig.3). In total, only 23 teams played more than 100 games in the EuroLeague between seasons 2000-2001 and 2018-2019. Only 17 of them played more than 100 games where the play-by-play data was recorded, between seasons 2007-2008 and 2018-2019. This is as much games as some NBA teams play in 1 season and gives us less data to work with for the prediction. models.

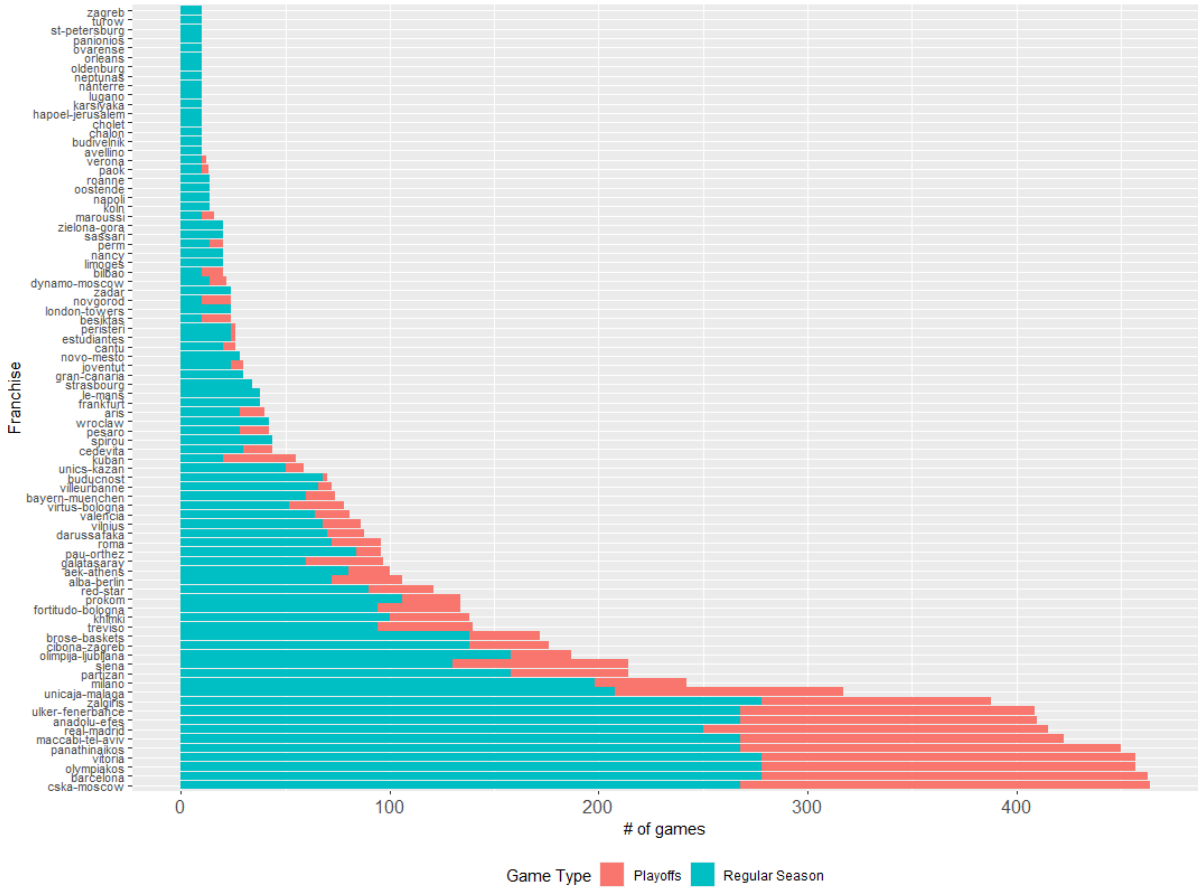


Figure 3. Games played per team in the EuroLeague from season 2000-2001 up to and including season 2018-2019. The blue bars show the number of regular season games and the red bars the number of playoff games played.

4.1.2 Correlations

Looking at the correlations between the different plays and the other variables in the seasonal average data, we find that some stronger ($\rho > 0.5$) but mostly weak ($\rho < 0.5$) correlations exist with multiple variables, and that all correlations are linear. The strongest correlation for the TP, 0PP, and 2PP was with the number of overtimes.

This is not surprising, because adding minutes to the game will cause more plays to be made (TP), and 0P and 2PP are the most common plays in a game. The number of overtimes a game will have is, however, not known beforehand. Some attempts were made to predict any overtimes, but they were not successful. The 4PP had the weakest correlations with any variable of the 5 plays, with its highest being $\rho = 0.12$ in the NBA and $\rho = 0.07$ in the EuroLeague. No good regression models are expected for the number of 4PP. The 3PP had the strongest correlation both in the NBA ($\rho = 0.60$) and EuroLeague ($\rho = 0.28$), with three-point related variables. The strongest correlation values for the TP, 0PP, 1PP, and 2PP lay between these values in both leagues. Prediction models for the NBA are expected to be more accurate than for the EuroLeague.

4.1.3 Total Plays (TP)

Figures 4 and 5 show the boxplots for the total number of plays per game, per season for the NBA and EuroLeague respectively. The year in the x-axis indicates the year the season ended. Individual games with overtimes are additionally plotted as points, except for games with one overtime for the NBA to avoid cluttering. The boxplots seem to be slightly positively skewed, as is expected for Poisson distributed data. The number of total plays in the NBA range from 85 to 160, while in the EuroLeague they range from 64 to 111, which is no surprise as NBA games are longer than EuroLeague games (twelve minutes per quarter vs ten minutes per quarter). As expected intuitively, games with overtimes are associated with a higher number of plays in both NBA and EuroLeague. A slightly increasing trend is visible in the total number of plays in the NBA, starting around the season 2011-2012. No clear trend is visible for the EuroLeague. This rising trend for the NBA is also visible in figure 6, which shows the mean number of total plays per minute for both leagues. An especially big jump was made from season 2017-2018 to season 2018-2019. The pace in the NBA is always higher than in the EuroLeague, as seen by the higher average total number of plays, of about 0.2 plays per minute (Fig.6).

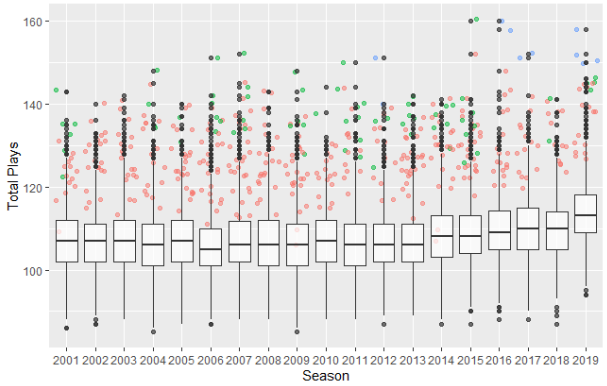


Figure 4. Boxplots for the total number of plays per game, per season in the NBA. The individual points depict the total plays of games with 2, 3, or 4 overtimes, in red, green and blue respectively.

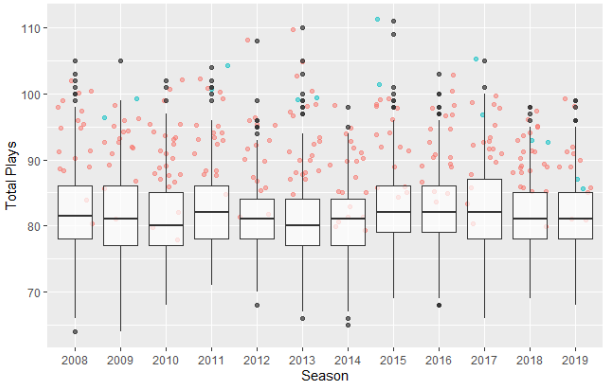


Figure 5. Boxplots for the total number of plays per game, per season in the EuroLeague. The individual points depict the total plays of games with 1 or 2 overtimes in red and blue respectively.

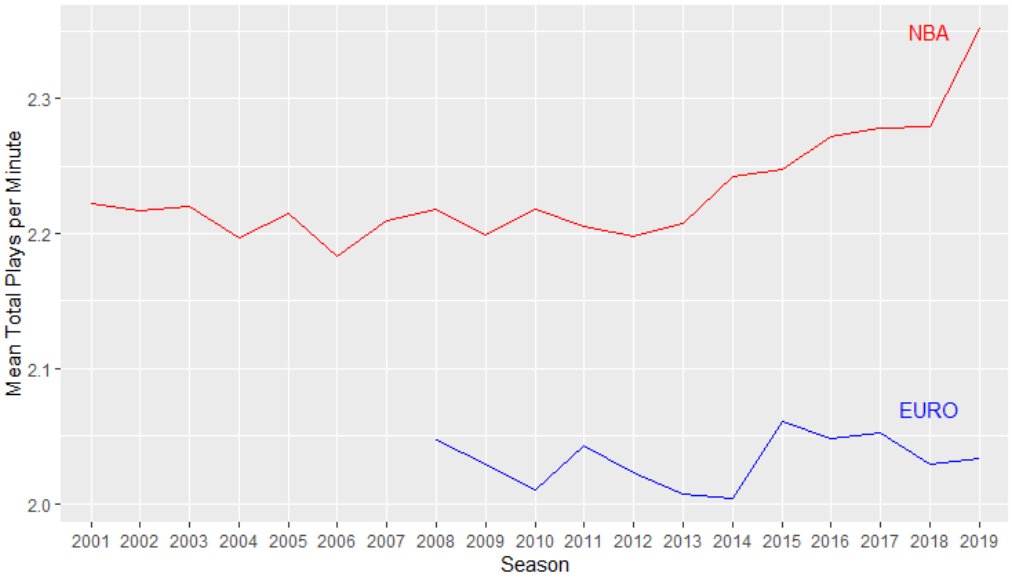


Figure 6. Mean total plays per minute, per season for both NBA (red) and EuroLeague (blue).

4.1.4 Zero-Point Plays (0PP)

The same plots are plotted as before but for the number of 0PP. The red line in figure 7 and figure 8 represents half of the average total number of plays per game, per season, just as a reference. This shows that more than 50% of the plays in a game result in zero points. Again, the association between the number of overtimes and the number of 0PP is visible but it is less pronounced as for the TP. The number of 0PP follows a very similar trend as the TP. This is expected as more than 50% of the total number of plays are 0PP. This is also noticeable when comparing figure 6 and figure 9. For the NBA we see a decline in 0PP from season 2003-2004 until season 2005-2006. This decline was not as clearly present in the TP.

The number of 0PP in the NBA then starts to slowly rise from season 2010-2011 until the last recorded season, where the jump in number of 0PP from season 2017-2018 to season 2018-2019 is also visible. The EuroLeague does not show a clear trend in 0PP but it does show a small dip in the number of 0PP in the last years, in contrast to the NBA. The difference in average 0PP per minute between both leagues is around 0.1 plays per minute.

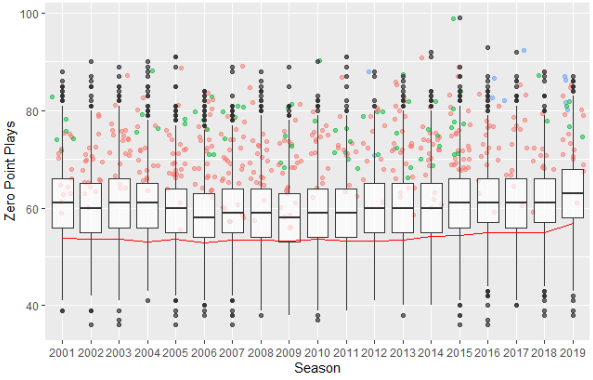


Figure 7. Boxplots for the total number of zero point plays per game, per season in the NBA. The individual points depict the total plays of games with 2, 3, or 4 overtimes in red, green and blue respectively. The red line represents half of the mean total plays per game played per season in the NBA as a reference line.

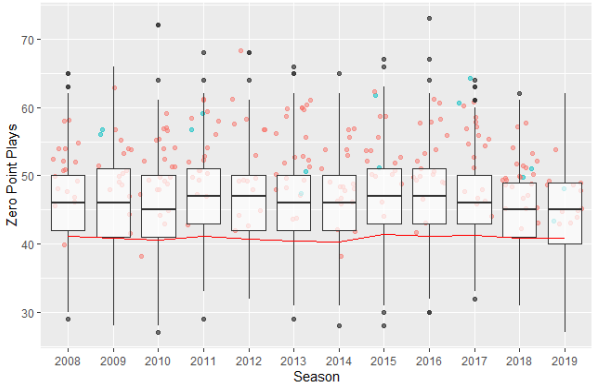


Figure 8. Boxplots for the total number of zero point plays per game, per season in the EuroLeague. The individual points depict the total plays of games with 1 or 2 overtimes in red and blue respectively. The red line represents half of the mean total plays per game played per season in the EuroLeague as a reference line.

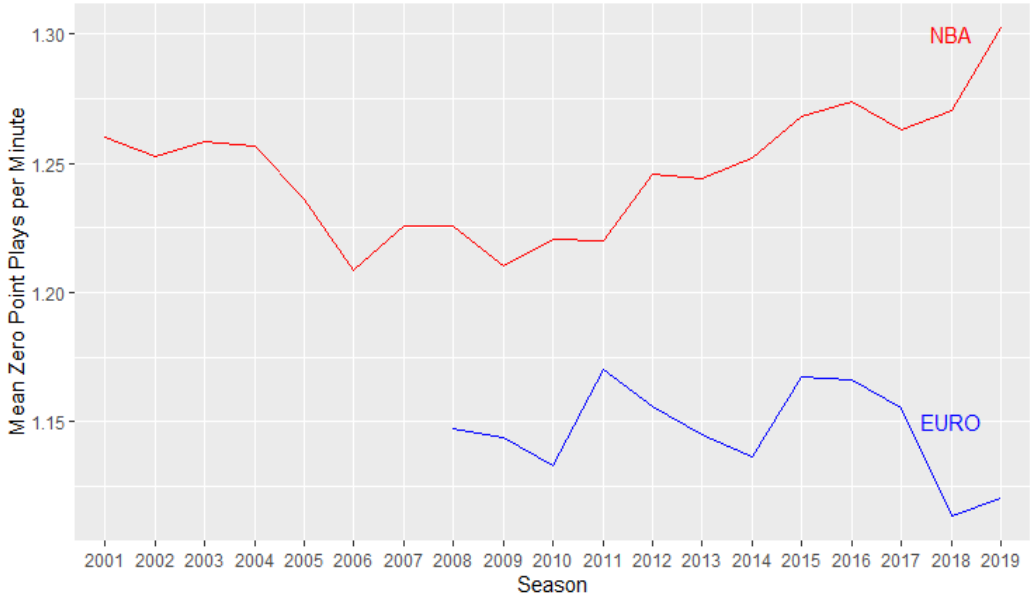


Figure 9. Mean number of zero point plays per minute, per season for both NBA (red) and EuroLeague (blue).

4.1.5 One-Point Plays (1PP)

Figures 10 and 11 show no clear association between the number of overtimes and the number of 1PPs, just as found by the correlation analysis. The number of 1PPs range from 0 to 17 in the NBA and only from 0 to 11 in the EuroLeague. The average 1PP per minute is lower in the NBA than in the EuroLeague in most seasons, and equal in three seasons but never higher (Fig.12). The difference, however, is very small, of about 0.018 1PP per minute. There is a slight increase in 1PP in the season 2005-2006 in the NBA (Fig.12). After the season 2010-2011, the number of 1PP seems to decline. Figure 12 also shows a small decline in 1PP after the season 2010-2011 until the season 2013-2014 in the EuroLeague.

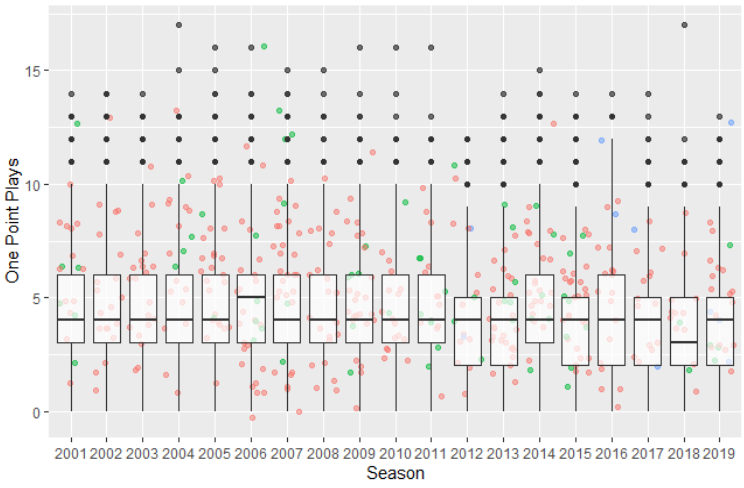


Figure 10. Boxplots for the number of one-point plays per game, per season in the NBA. The individual points depict the total plays of games with 2, 3, or 4 overtimes in red, green and blue respectively.

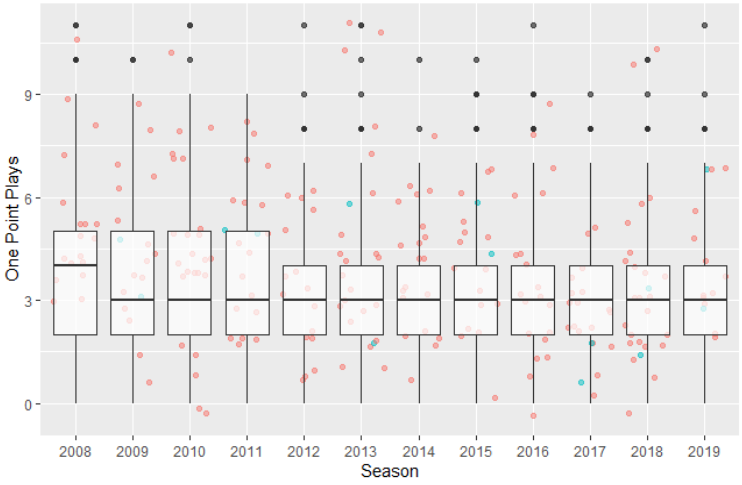


Figure 11. Boxplots for the number of one-point plays per game, per season in the EuroLeague. The individual points depict the total plays of games with 1 and 2 overtimes in red and blue respectively.

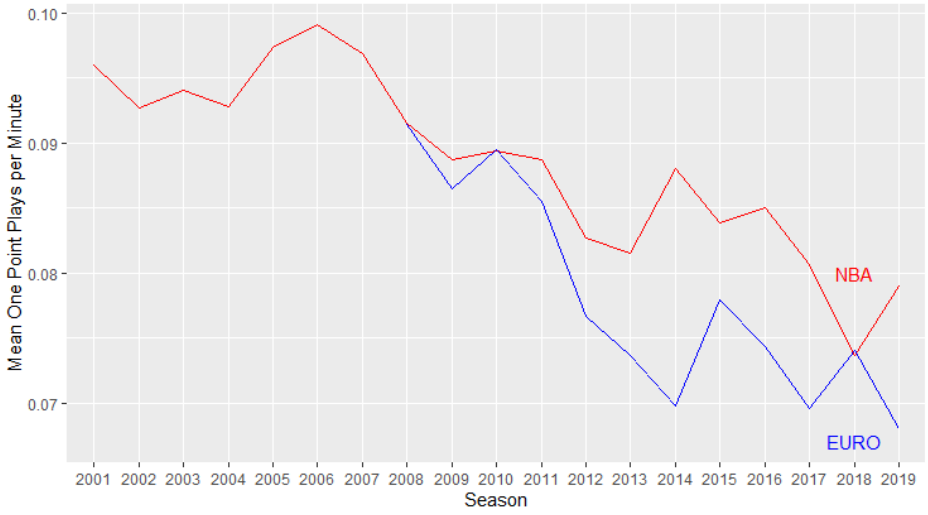


Figure 12. Mean number of one-point plays per minute, per season for both NBA (red) and EuroLeague (blue).

4.1.6 Two-Point Plays (2PP)

The seasonal boxplots for the NBA show once again an association between the number of 2PP and the number of overtimes, but it is less pronounced than for the 0PP data. The number of 2PP in the NBA range from as little as 14 to as much as 62 (Fig. 13.), and in the EuroLeague from 9 to 42 (Fig. 14). On first sight there is also no upward trend in the number of two-points possessions over time, it stays more or less equal, with maybe a small decline from season 2009-2010 until season 2017-2018 in the NBA (Fig 15). This decline is not present in the EuroLeague, where the mean 2PP has remained approximately the same over the seasons. Again, the average number of 2PP per minute is noticeably lower for the EuroLeague compared to the NBA, with more than 0.1 2PP per minute.

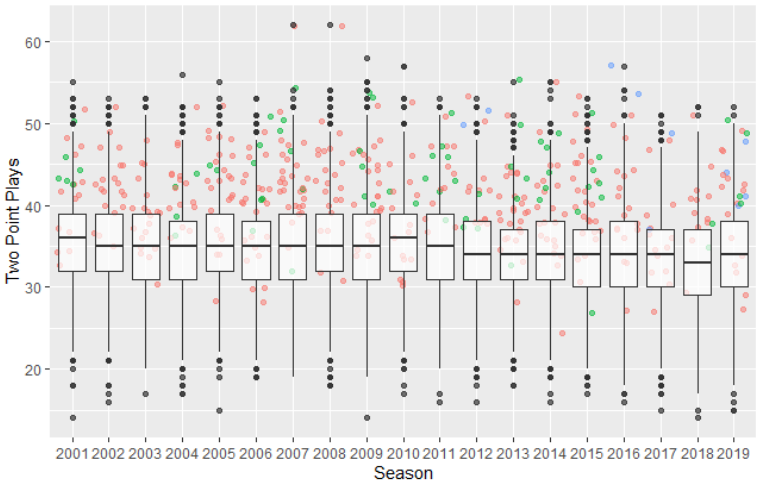


Figure 13. Boxplots for the number of two-point plays per game, per season in the NBA. The individual points depict the total plays of games with 2, 3, or 4 overtimes in red, green and blue respectively.

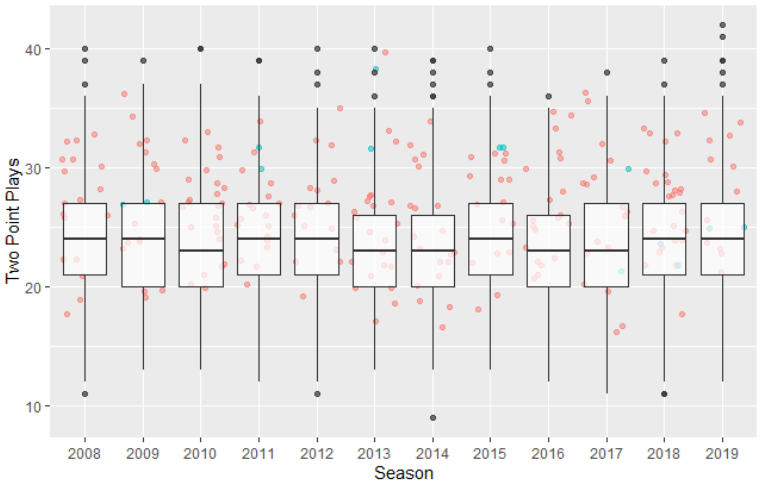


Figure 14. Boxplots for the number of two-point plays per game, per season in the EuroLeague. The individual points depict the total plays of games with 1 and 2 overtimes in red and blue respectively.

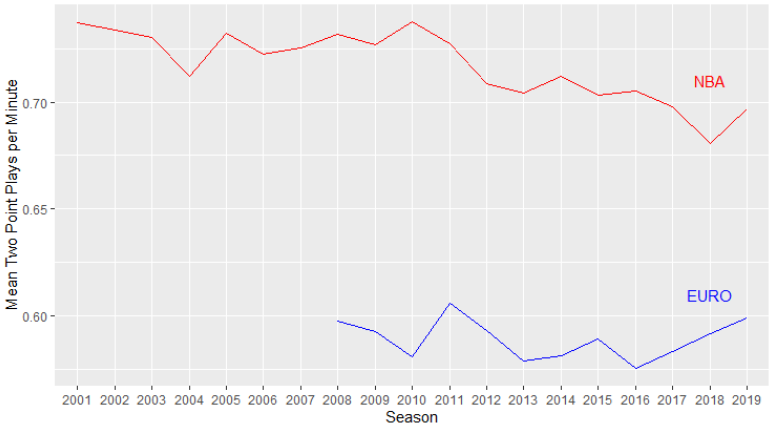


Figure 15. Mean number of two-point plays per minute, per season for both NBA (red) and EuroLeague (blue).

4.1.7 Three-Point Plays (3PP)

There is no clear association visible between the number of overtimes and the number of 3PP in both the NBA and the EuroLeague. There is, however, a clearly increasing trend in number of 3PP in the NBA (Fig. 16 & 18). The number of 3PP per game doubles between season 2000-2001 and season 2018-2019, from around 6 on average to around 12. This rising trend is also visible for the EuroLeague, although less pronounced (Fig.17 & 18). The number of 3PP per minute has been increasing constantly for the NBA, with a slight decrease between seasons 2008-2009 and 2011-2012. The EuroLeague also shows an increasing trend, but with a sharp drop in mean 3PP per min in the season 2010-2011. Very noticeable is that, unlike for the other types of plays, the mean number of 3PP per minute in the EuroLeague was higher than in the NBA for all the recorded seasons, except the last two seasons. This difference, however, is relatively small, of about 0.04 mean 3PP per minute.

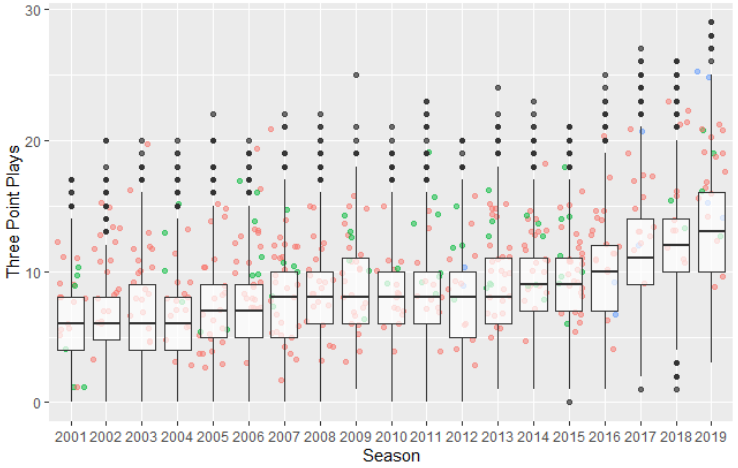


Figure 16. Boxplots for the number of three-point plays per game, per season in the NBA. The individual points depict the total plays of games with 2, 3, or 4 overtimes in red, green and blue respectively.

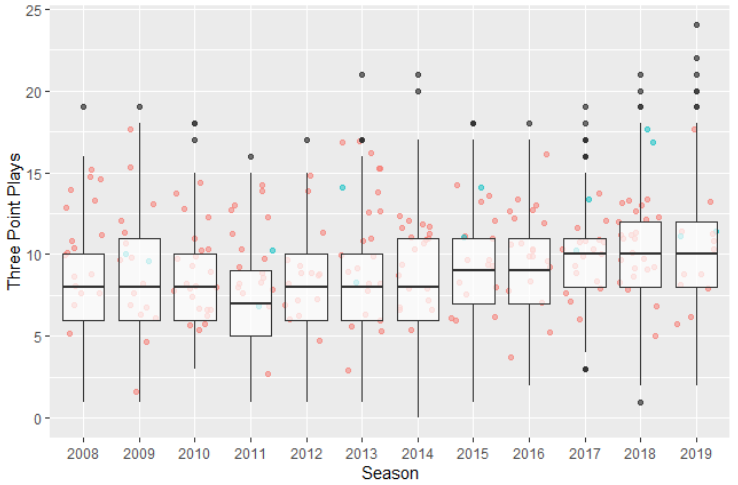


Figure 17. Boxplots for the number of three-point plays per game, per season in the EuroLeague. The individual points depict the total plays of games with 1 and 2 overtimes in red and blue respectively.

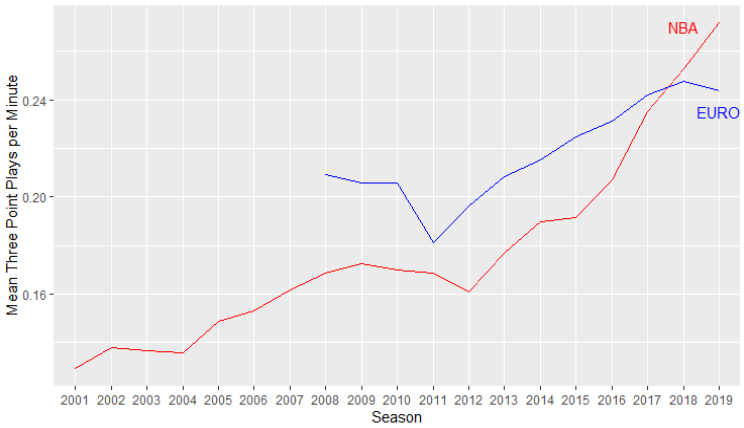


Figure 18. Mean number of three-point plays per minute, per season for both NBA (red) and EuroLeague (blue).

4.1.8 Four-Point Plays (4PP)

A 4PP is a rare event that only happens when a three-point shot is made and fouled, and the subsequent free-throw scored. Intuitively, the number of 4PP is thus dependent on the number of three-point shots made and attempted. A strong increasing number of 4PP per season in both leagues follows the increase in number of 3PP (Fig 16-21). The rarity of the event is also apparent from the fact that of the approximately 1310 games played each season in the NBA, less than 50 had a single 4PP in the first recorded seasons, while only 250 had a 4PP in the season with the highest number of 4PP. The same goes for the EuroLeague, where of the 188-253 games played per season in the EuroLeague, the first seasons had less than ten games with a single 4PP and the season with the highest amount of 4PP had 38 games with 4PP. The increasing trend is also shown in the mean 4PP per minute, which is approximately the same in both leagues (Fig 21).

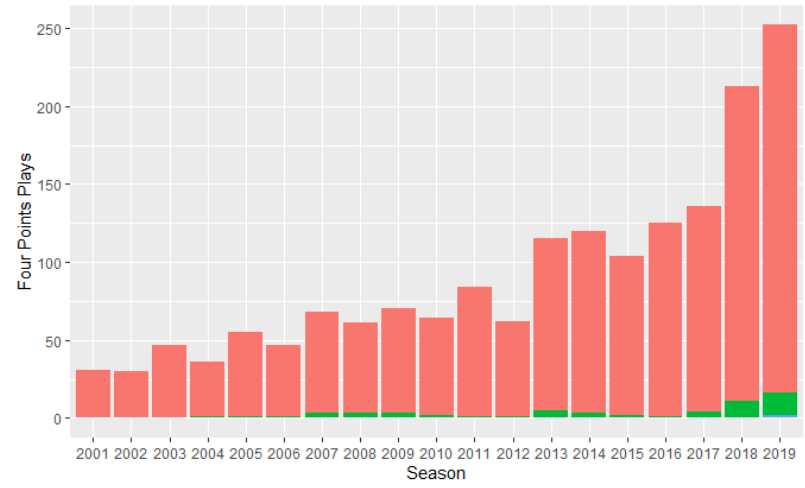


Figure 19. Total number of games with four-point plays per season in the NBA, with in red, green, and blue the number of games with one, two, and three four point plays respectively.

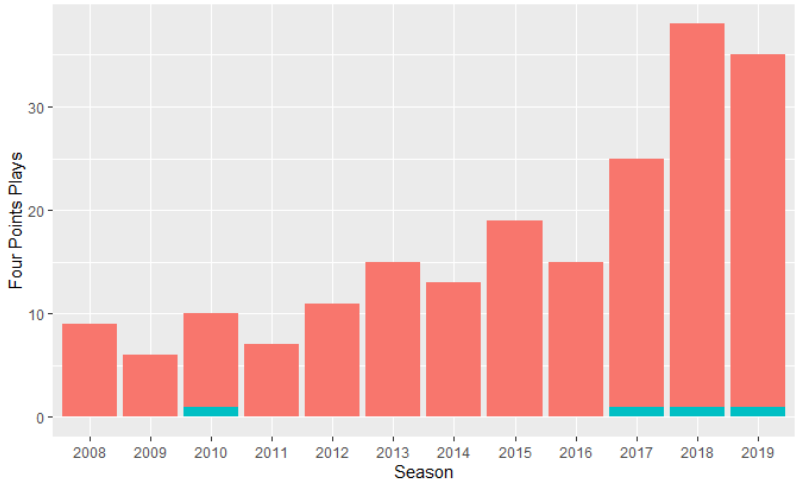


Figure 20. Total number of games with four point plays per season in the EuroLeague, with in red and blue the number of games with one and two four point plays respectively.

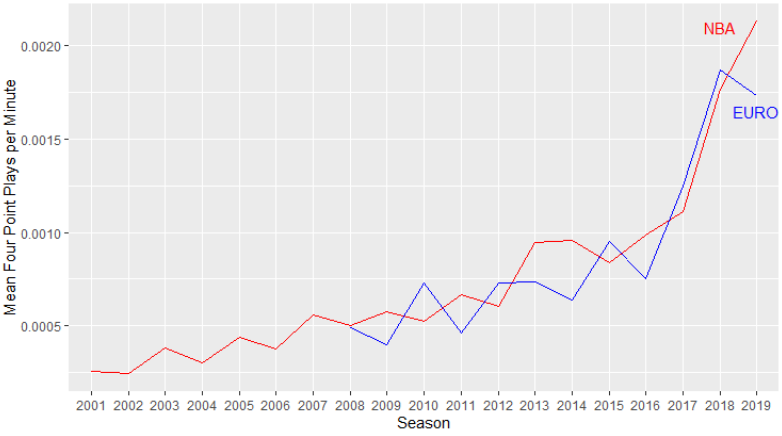


Figure 21. Mean number of four-point plays per minute, per season for both NBA (red) and EuroLeague (blue).

4.2 Predicting Plays

4.2.1 General model

	RMSE	Deviance	Rsquare
TP Training	6.4741112	0.3834079	0.2837785
TP Test	6.5655525	0.3938280	0.2832498
0PP Training	6.6930427	0.7420202	0.1460382
0PP Test	6.7018672	0.7445354	0.1505397
1PP Training	2.1682284	1.1617384	0.0719983
1PP Test	2.1763477	1.1637882	0.0684922
2PP Training	4.9283082	0.7066849	0.1830687
2PP Test	4.9720717	0.7189536	0.1783090
3PP Training	2.9880070	1.0707440	0.3752344
3PP Test	3.0345349	1.1197982	0.3603928
4PP Training	0.1906152	0.2228028	0.0809029
4PP Test	0.1952267	0.2368098	0.0557878

Table 3. Evaluation results of the NBA models predicting the TP, 0PP, 1PP, 2PP, 3PP, and 4PP on the training set and the test set, using the Root Mean-Squared Error (RMSE), deviance R-squared, and Deviance.

	RMSE	Deviance	Rsquare
TP Training	5.4380442	0.3584601	0.1668476
TP Test	5.5300932	0.3713185	0.1304859
0PP Training	5.6385012	0.6922122	0.1593794
0PP Test	5.8636802	0.7413241	0.1357915
1PP Training	1.7949589	1.1272899	0.0797135
1PP Test	1.8647364	1.2082959	0.0589373
2PP Training	4.3679250	0.8090169	0.1611140
2PP Test	4.2146881	0.7498658	0.1221810
3PP Training	2.8898299	0.9687730	0.1623193
3PP Test	2.8704525	0.9742694	0.1309007
4PP Training	0.1891982	0.2224276	0.0887273
4PP Test	0.2065492	0.2717024	-0.0020432

Table 4. Evaluation results of the EuroLeague models predicting the TP, 0PP, 1PP, 2PP, 3PP, and 4PP on the training set and the test set, using the Root Mean-Squared Error (RMSE), deviance R-squared, and Deviance.

Tables 3 and 4 show the RMSE, Deviance and R^2_{DEV} for the different prediction models for the NBA and EuroLeague respectively. Both tell a similar story. There is only a small difference in RMSE and Deviance between training and test set. This means there is little to no overfitting in the models. Looking to the R^2_{DEV}, the prediction models for the TP, 0PP, 2PP, and 3PP have the highest R^2_{DEV} and are thus the models that will give the best predictions. The R^2_{DEV} for the 1PP and 4PP are very low, smaller than 0.09, meaning the models will give very poor predictions, only very slightly better than predicting the league average. The R^2_{DEV} for the other kind of plays, however, are not very high either. The R^2_{DEV} of the model predicting the 3PP was the highest, with $R^2_{DEV} = 0.36$ for the NBA and $R^2_{DEV} = 0.16$ for the EuroLeague, as was expected from the correlation analysis. Difficulty in getting accurate predictions for the different kinds of plays are logical, due to their inherent randomness. That said, even small improvements in prediction over the average introduce differences between teams that could be useful in the simulations later on. For example, two teams averaging the league average in 0PP, 1PP, 3PP and 4PP per game play against each other but one team averages one 2PP more per game. Simulating this game thousands of times will favour the team averaging one more 2PP over the other.

Many of the significant variables are related to the season and the teams playing, next to multiple box score statistics of both the team and its opponent. See Appendix 5 for the significant variables in the models.

4.2.2 Team Models

4.2.2.1 No Interactions

Only the model without interactions is discussed in detail because it is the model with the highest accuracy in simulating game outcomes in both leagues.

4.2.2.2 Evaluation Measures

The RMSE and the deviance of the individual NBA team models give approximately the same values as the RMSE and deviance of the model discussed above (Table 3). Their differences between teams is generally relatively small, showing that the models for the different teams have a similar performance (Appendix 6). There is some variation in the R^2_{DEV} between teams but the R^2_{DEV} for all the different plays was generally greater than zero, except for the 4PP and 1PP where many were zero. In those cases, the average was thus used as predicted value. The general trends seen before however did still hold. The models with the best absolute fit in terms of R^2_{DEV}, were the models for the 3PP, followed by the models for the TP, 2PP, 0PP, 1PP, and 4PP in that order.

The results for the EuroLeague tell a different story (Appendix 6). The R^2_{DEV} of the models applied on the training sets varied between zero and one, but most R^2_{DEV} of the models applied on the test set were zero. This indicates heavy overfitting. The teams with the worse fits are generally the teams that only played one season in the EuroLeague and did not have much data. Their training and test set comprises of games out of a single season. The teams with the better model fits were then inversely the teams that played multiple seasons in the EuroLeague. The models with the least overfitting are the models of the teams with the most seasons played in the EuroLeague (Fig. 3). Only 18 teams played more than 82 matches in the EuroLeague between seasons 2007-2008 and 2017-2018, which is how many games an NBA team plays in one season.

4.2.2.3 Total Plays (TP)

Firstly, the difference in the number of models that share the most common variables in each league is noticeable (Table 6 & 7).

In the NBA, each team model (all 30) predicts a positive effect of the average number of TP of the opponent (Opp_TotalPoss) on the number of total plays of the team. In the EuroLeague, it only had a positive effect on 15 of the 61 teams. The list of shared variables between teams within one league is also shorter for the EuroLeague.

The opponent seems to play an important role in predicting the number of TP a team makes in a game. In the NBA, playing against fast paced teams (high average Opp_TotalPoss) such as the Golden State Warriors, the Phoenix Suns and the Sacramento Kings, has a positive influence on the number of TP of a 29, 23, and 22 of the teams respectively. Playing against slow paced teams (low average Opp_TotalPoss) such as the Utah Jazz, San Antonio Spurs and Detroit Pistons had an opposite effect for 28, 24 and 21 of the teams respectively. The same happens in the EuroLeague, where Maccabi Tel Aviv had a positive influence on the number of TP of 12 teams. Also, which season the game is played in plays an important role in many of the NBA models, while it does not in the EuroLeague.

Playing against teams that concedes the ball a lot, score a lot of points and blocks a lot of shots will boost the team's number of TP in the NBA. In the EuroLeague, the average number of field goal attempts and steals of the opponent are the most important positive variables. Not important for all teams but playing at home also results in more TP in both leagues for some. Playing against a team that takes a lot of offensive rebounds will result in less TP for the team in the NBA. Playing against a team that wins a lot and playing a playoff game (GameType/game_type equal to one) results in less TP in both the NBA and the EuroLeague.

Variable	Count
Opp_TotalPoss	30
OpponentGSW	29
Opp_TOV	25
Opp_PTS	24
OpponentPHO	23
OpponentSAC	22
Opp_BLK	21
Opp_FG	18
Opp_Losses	18
Opp_AST	17
OpponentDEN	16
Opp_STL	15
Season2003	12
Season2005	12
OpponentBOS	12
OpponentIND	12
OpponentPHI	12
Season2004	11
Team_TeamPoss	11
HomeAdvantage	11
OpponentCHI	10
OpponentORL	10
Opp_FGA	10
Opp_2P%	10
Opp_3PA	10

Variable	Count
Opp_TotalPoss	15
Opponentmaccabi-tel-aviv	10
Opp_FGA	8
Opp_STL	7
HomeadvantageTRUE	7
Opp_PTS	6
Opp_1PtsPoss	6
Opp_PF	5

Table 6. Number of times a variable had a significant positive effect on the number of total plays (TP) in all the team models in the NBA (left) and EuroLeague (right/above) . Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

Variable	Count
Opp_ORB	28
OpponentUTA	25
OpponentSAS	24
game_type	19
OpponentDET	19
OpponentNOH	19
OpponentPOR	17
OpponentCLE	16
Opp_Wins	15
Season2009	14
OpponentMEM	13
Season2007	12
Season2008	12
Season2011	12
OpponentMIA	12
Season2006	11
Season2014	11
Team_DRB%	11
OpponentMIN	11
Opp_TRB	11
Season2013	10
OpponentTOR	10
Opp_3P%	10
Team_4PtsPoss	10
Team_ConsRoadGames	10

Variable	Count
Team_Wins	8
GameType	7
Team_DRB%	6
Team_BLK%	5
Opponentmilano	5
Opponentulker-fenerbahce	5

Table 7. Number of times a variable had a significant negative effect on the number of total plays (TP) in all the team models in the NBA (left) and EuroLeague (above). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

4.2.2.4 Zero-Points Plays (0PP)

Just as for the number of TP, playing against fast paced teams results in a higher number of 0PP in all but one NBA teams and in 6 EuroLeague teams (Table 8). Furthermore, if the opponent averages high numbers in its defensive statistics, such as steals, defensive rebounds, and blocks, the team will have more 0PP. In the EuroLeague this is also the case but for many more teams a high average number of blocks of the opponent seems to be important than a high average number of defensive rebounds or steals. Maybe more surprising is that the opponent's average number of assists and turnovers also has a positive impact on the teams 0PP in the NBA, and the opponent's average number of assists and field goal attempts in the EuroLeague. Here again, playing against fast-paced teams such as the Golden State Warriors, the Phoenix Suns, or Maccabi Tel Aviv results into more 0PP for many teams. For the EuroLeague, the number of consecutive road games a team is playing has a positive impact on the 0PP in many EuroLeague teams.

This influence is strengthened by the fact that playing at home is the only variable with a negative influence on the number of 0PP of eleven teams in the EuroLeague (Table 9). Home advantage is also the most common negative influence on the number of 0PP in NBA teams followed by playing against teams that lost many games. This is reinforced for some teams who already won many games (Team_Wins). For the NBA, similar to the number of TP, playing against slower paced opponents or less defensive teams such as the Utah Jazz, Cleveland Cavaliers, Portland Trailblazers, and Toronto Raptors results in less 0PP for many teams.

Variable	Count
Opp_TotalPoss	29
Opp_STL	28
Opp_DRB	27
Opp_BLK	25
Opp_AST	23
Opp_TOV	20
OpponentGSW	17
OpponentPHO	17
Opp_PTS	17
OpponentIND	13
Team_0PtsPoss	13
Season2012	11
Season2015	11
Team_TOV	11
OpponentPHI	11
OpponentSAC	11
Season2001	10
Season2003	10
Season2004	10
Season2005	10
OpponentDEN	10
Opp_2P%	10

Variable	Count
Opp_BLK	10
Opponentmaccabi-tel-aviv	7
Opp_FGA	7
Opp_AST	7
Team_ConsRoadGames	7
Opp_FG	6
Opp_TotalPoss	6
Season2015	5
Opp_DRB	5
Opp_STL	5
Opp_BLK%	5

Table 8. Number of times a variable had a significant positive effect on the number of zero point plays (0PP) in all the team models in the NBA (left) and EuroLeague (right/above). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

Variable	Count
HomeAdvantage	25
Opp_Losses	23
OpponentUTA	19
Season2011	14
OpponentPOR	13
OpponentTOR	13
Team_Wins	13
Season2008	12
Season2009	12
Team_FT%	12
OpponentDET	12
OpponentNYK	12
OpponentCLE	11
OpponentNOH	11
Opp_ORB	10

Variable	Count
HomeadvantageTRUE	11

Table 9. Number of times a variable had a significant negative effect on the number of zero point plays (0PP) in all the team models in the NBA (left) and EuroLeague (right/above). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

4.2.2.5 One-Point Plays (1PP)

As mentioned before, the number of one-point plays are more difficult to predict because their number is generally low, sometimes even zero. In both leagues, playing against an aggressive team averaging a lot of personal fouls will result in more 1PP (Table 10). This expected as a foul needs to be committed to have a chance at a 1PP. Also playing at home results in more 1PP in both leagues. Further the average number of free-throw attempts and free-throw rate of the team as well as the game type were important in the NBA. In the EuroLeague, the teams average effective field goal percentage, and the opponent's average number of turnovers and block percentage resulted in a higher number of 1PP.

Playing against the San Antonio Spurs had a negative effect on the number of 1PP for 15 NBA teams, while season 2017-2018 seems to be a year with less 1PP for 14 of them (Table 11). Additionally, a high average free-throw percentage of the team lowered the number of predicted 1PP. In the EuroLeague, the positive impact of home advantage on the predicted number of 1PP is again accentuated by the negative impact of consecutive road games. Playing against a strong team with a lot of wins is detrimental to the team's number of 1PP, while the team's number of wins and losses, as well as its average number of free throw (attempts) also resulted in less 1PP in the EuroLeague.

Variable	Count
Opp_PF	30
Team_FTA	21
Team_FTr	20
game_type	14
HomeAdvantage	14
Team_1PtsPoss	12
Team_ORB%	11
Season2006	10

Variable	Count
Opp_PF	14
HomeadvantageTRUE	8
Team_eFG%	6
Opponentroma	6
Opp_TOV	6
Opp_BLK%	6
Team_FG%	5
Team_3P%	5
Team_DRB%	5
Opponentmaccabi-tel-aviv	5
Opponentpartizan	5
Opp_3PAr	5
Opp_Losses	5

Table 10. Number of times a variable had a significant positive effect on the number of one-point plays (1PP) in all the team models in the NBA (left) and EuroLeague (right). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

Variable	Count
OpponentSAS	15
Season2018	14
Team_FT%	11

Variable	Count
Opp_Wins	8
Team_FTA	7
Opponentbarcelona	7
Team_Wins	7
Team_ConsRoadGames	7
Team_FT	6
Team_Losses	6
Opponentvitoria	5
Opp_FTA	5
Opp_TS%	5

Table 11. Number of times a variable had a significant negative effect on the number of one-point plays (1PP) in all the team models in the NBA (left) and EuroLeague (right). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

4.2.2.6 Two-Point Plays (2PP)

Playing at home caused 29 NBA teams and 9 EuroLeague teams to have more 2PP than playing on the road (Table 12). Similar to the number of TP and 0PP, playing against fast paced teams (Opp_TotalPoss) will result in more 2PP in many teams. This is strengthened by the fact that having teams such as the Golden State Warriors and the Phoenix Suns in the prediction model 16 and 17 NBA teams. As one would expect, a team that attempts a lot of two-point field goals will get more 2PP but also playing against teams that average a high number of blocks and points will up that number in the NBA. Important in all NBA teams, however, is playing against a bad team, with a lot of lost games. In the EuroLeague, having Milano as opponent, as well as playing against a team with a high average number of 0PP, personal fouls and points of the opponent was important.

The average number of steals of the opponent was the most common detrimental to the number of 2PP in the NBA, coincidently also being one of the most common positive variables to predict the number of 0PP in the NBA (Table 8), and also had a negative impact in the EuroLeague. Game type was the second most negatively impactful variable in both leagues and has a similar influence as for the TP prediction models (Table 7). Teams tend to play slower in the playoffs compared to the regular season and also have less 2PP. Again, the negative impact of the number of consecutive road games strengthens the positive influence of home advantage in the EuroLeague, as seen for the 0PP (inverse relation) and the 1PP. Other important variables in the NBA are the number of wins and the average block percentage of the opponent, and the average three-point attempt (rate) of the team.

Variable	Count
Opp_Losses	30
Opp_TotalPoss	29
HomeAdvantage	29
Team_2PA	19
OpponentPHO	17
Opp_BLK	17
Opp_PTS	17
OpponentGSW	16
OpponentTOR	16
Season2010	15
Opp_FG	15
Season2008	14
Season2003	13
Season2009	13
OpponentSAC	13
Opp_FGA	13
Opp_TOV	12
Season2002	11
Season2005	11
Team_2P	11
Team_AST%	11
Opp_FT%	11
Season2001	10
Season2006	10
Season2007	10
Team_FT	10
OpponentMIL	10

Variable	Count
HomeadvantageTRUE	9
Opponentmilano	6
Opp_PF	6
Opp_PTS	6
Opp_0PtsPoss	6
Opp_TotalPoss	6
Opponentolimpija-ljubljana	5
Opp_FT%	5

Table 12. Number of times a variable had a significant positive effect on the number of two-point plays (2PP) in all the team models in the NBA (left) and EuroLeague (right/above). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

Variable	Count
Opp_STL	24
game_type	22
Opp_BLK%	20
Opp_Wins	20
Team_3PAr	19
OpponentMIA	19
OpponentMEM	16
OpponentUTA	16
Team_3PA	15
Opp_TRB	15
Opp_PF	15
Opp_AST%	15
OpponentCHA	14
Season2004	13
Season2018	13
Opp_DRB	13
Season2012	12
OpponentBOS	12
OpponentNOH	12
Team_4PtsPoss	12
Season2013	11
Season2015	11
Season2016	11
Season2017	11
OpponentPHI	11
Season2014	10
Team_3P	10
OpponentOKC	10
Opp_4PtsPoss	10
Team_ConsRoadGames	10

Variable	Count
Opponentpanathinaikos	8
Team_ConsRoadGames	8
GameType	7
Opponentcska-moscow	6
Opp_STL	6
Opponentmaccabi-tel-aviv	5

Table 13. Number of times a variable had a significant negative effect on the number of two-point plays (2PP) in all the team models in the NBA (left) and EuroLeague (right/above). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

4.2.2.7 Three-Point Plays (3PP)

As expected from the correlation analysis, three-point related statistics of the team were the most common variables used to predict the number of 3PP of the team (Table 14). In the NBA, playing against more fast-paced teams, and bad teams that lost a large amount of games already also facilitated the number of 3PP of the team. Playing against an opponent that attempts a lot of three-point shots also resulted in more 3PP by the team in both leagues. Other factors resulting in more 3PP in the EuroLeague were playing against teams that like to steal the ball, home advantage and getting a lot of offensive rebounds.

For the first time we see that a season pops up in the common variables for the EuroLeague. For five teams, Season 2010-2011 saw a decrease in 3PP in the EuroLeague (Table 15). These trends can also be found in figure 18. Other variables also reduce the number of 3PP, such as the game type or playing consecutive road games, but only in maximum five EuroLeague teams (Table 15). In the NBA, playing against defensive teams, such as the San Antonio Spurs and the Detroit Pistons will result in less 3PP, as will playing against teams with on average a high number of blocks and offensive rebounds. Teams that attempt on average a lot of two-point field goals and have on average a high offensive rebound percentage will also make less 3PP.

Variable	Count	Variable	Count
Team_3PA	30	Team_3PA	8
Team_3PAr	30	Opp_3PA	7
Team_3P	27	Opp_STL	6
Team_3PtsPoss	27	Opp_4PtsPoss	6
Opp_TotalPoss	21	Team_3PAr	5
Opp_Losses	20	`Team_ORB%`	5
Opp_FGA	18	HomeadvantageTRUE	5
Team_PTS	16		
Opp_3PA	15		
Opp_PTS	15		
Season2018	14		
OpponentNYK	14		
HomeAdvantage	14		
Team_FGA	13		
Season2014	11		
Season2017	11		
Team_TeamPoss	11		
Opp_3PtsPoss	11		
Season2009	10		
`Team_2P%`	10		
OpponentCHA	10		
OpponentDEN	10		
OpponentMIL	10		
OpponentWAS	10		
Team_4PtsPoss	10		

Table 14. Number of times a variable had a significant positive effect on the number of three-point plays (3PP) in all the team models in the NBA (left) and EuroLeague (right/above). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

Variable	Count	Variable	Count
OpponentSAS	28	Season2011	5
Opp_BLK	27	GameType	5
Team_2PA	18	`Team_3P%`	5
OpponentDET	17	Team_PTS	5
`Team_ORB%`	16	Opponentbarcelona	5
Opp_Wins	15	Opp_FTA	5
Season2001	14	Opp_PF	5
OpponentBOS	14	Team_2PtsPoss	5
Opp_ORB	14	Opp_Wins	5
Team_TOV	13	Team_ConsRoadGames	5
Team_PF	13		
`Opp_AST%`	13		
Season2002	12		
Season2004	12		
Opp_TRB	12		
game_type	11		
`Team_AST%`	11		
OpponentHOU	11		
`Team_TRB%`	10		
OpponentCHI	10		
OpponentIND	10		
Team_ConsRoadGames	10		

Table 15. Number of times a variable had a significant negative effect on the number of three-point plays (3PP) in all the team models in the NBA (left) and EuroLeague (right/above). Only the variables that were significant in 10 or more models in the NBA, and 5 or more models in the EuroLeague are shown.

4.2.2.8 Four-Point Plays (4PP)

Predicting four-point plays is very difficult due to the scarcity of the play. None of the models could somewhat accurately predict any 4PP and thus the variables of these models will not be discussed in more detail. The few models with a smaller deviance than the model with only the intercept relied primarily on team statistics related to the number of three points and 3PP (3P, 3PA, 3Par, 3PtsPoss). This is expected because of the reliance of 4PP on fouled three-point shots.

In general, there are always a few common variables for almost all 30 NBA teams. This is not the case for the EuroLeague. A variable is at most shared among the models of 15 of the 61 teams. This is not too surprising as only 17 teams played more than 100 games in the EuroLeague between seasons 2007-2008 and 2017-2018, 27 teams played more than 50 in that time period. NBA teams play more games in one season than 44 of the 61 EuroLeague teams play in 10 seasons. Therefore there is simply less data available for EuroLeague teams to make accurate models. Most models for EuroLeague teams with a small number of games have very few, if any, additional variables other than the intercept. Another trend to note is that the season did not appear as an important variable in EuroLeague models, except for the 3PP, while it was prevalent in NBA models. This trend was expected from the data exploration (Fig 6, 9, 12, 15, 18, 21). In the EuroLeague, the different kind of plays did not show any significant trends over the seasons, except the 3PP, while they did in the NBA. Also the number of seasons, and again the number of games, were smaller for the EuroLeague than for the NBA causing smaller changes between seasons to be statistically insignificant. Interesting is also both the similarities and dissimilarities in importance of different variables between the two leagues. This shows that, even though they are playing the same game with the same rules, the dynamics are different.

4.3 Game Simulations

	Accuracy 2000-2018	Accuracy 2018-2019
No Prediction Model	0.654	0.645
General Model	0.654	0.645
Team Models No Interactions 1	0.650	0.665
Team Models No Interactions 2	0.651	0.673
Team Models Interactions	0.633	0.664

	Accuracy 2000-2018	Accuracy 2018-2019
No Prediction Model	0.679	0.764
General Model	0.679	0.764
Team Models No Interactions 1	0.711	0.786
Team Models No Interactions 2	0.693	0.779
Team Models Interactions	0.679	0.764

Table 16. Prediction accuracy of the game winner through simulations using the predicted plays of different models for the NBA (Up), and the EuroLeague (down).

In both leagues, the plays predicted by the team models with no interactions included between season and opponents gave the best results for predicting winners in the season 2018-2019. For the EuroLeague, the group of team models that trained the α parameter of the elastic net gave the best result (Table 16), while for the NBA the group of team models with the α parameter of the elastic net set to 0.1 gave the best results (Table 16). However, using the average number of the different plays by each team in each season gave only a slightly lower accuracy in the NBA and the EuroLeague. Predicting the different kinds of plays for each team instead of taking their average only resulted in 21 and 3 more games correctly predicted on the 764 and 140 games played in our test season 2018-2019 in the NBA and EuroLeague respectively. This shows us that trying to predict the different kinds of plays for each team, each game, using a complex model is not necessary to get good results out of the simulations, but it can potentially be used to improve them. Better predictions models are needed to improve the results more significantly.

Generally, the prediction accuracy was higher in the EuroLeague, where we correctly predicted 1808 of the 2529 games (71.1%) in all seasons, and 110 of the 140 games (78.6%) in our test season 2018-2019. In the NBA, we correctly predicted 15767 of the 24203 games (65.1%) in all seasons, and 514 of the 764 games (67.3%) in our test season 2017-2018. The probability distributions of the scores of each team, in each game, are normal distributed and have relatively large confidence intervals.

The 95% confidence intervals for NBA games are on average 58.4 points wide, and for EuroLeague games 52.4 points wide. The outcome of all NBA and EuroLeague games played fell within the 95% confidence interval of the simulations in 99.2-99.5% of the cases, and the outcome of all the games played in the test season fell within the 95% confidence interval.

Prediction for individual games and the probability distribution of the scores of each team can be viewed using the basic shiny app attached to the paper (e.g. Fig. 22).

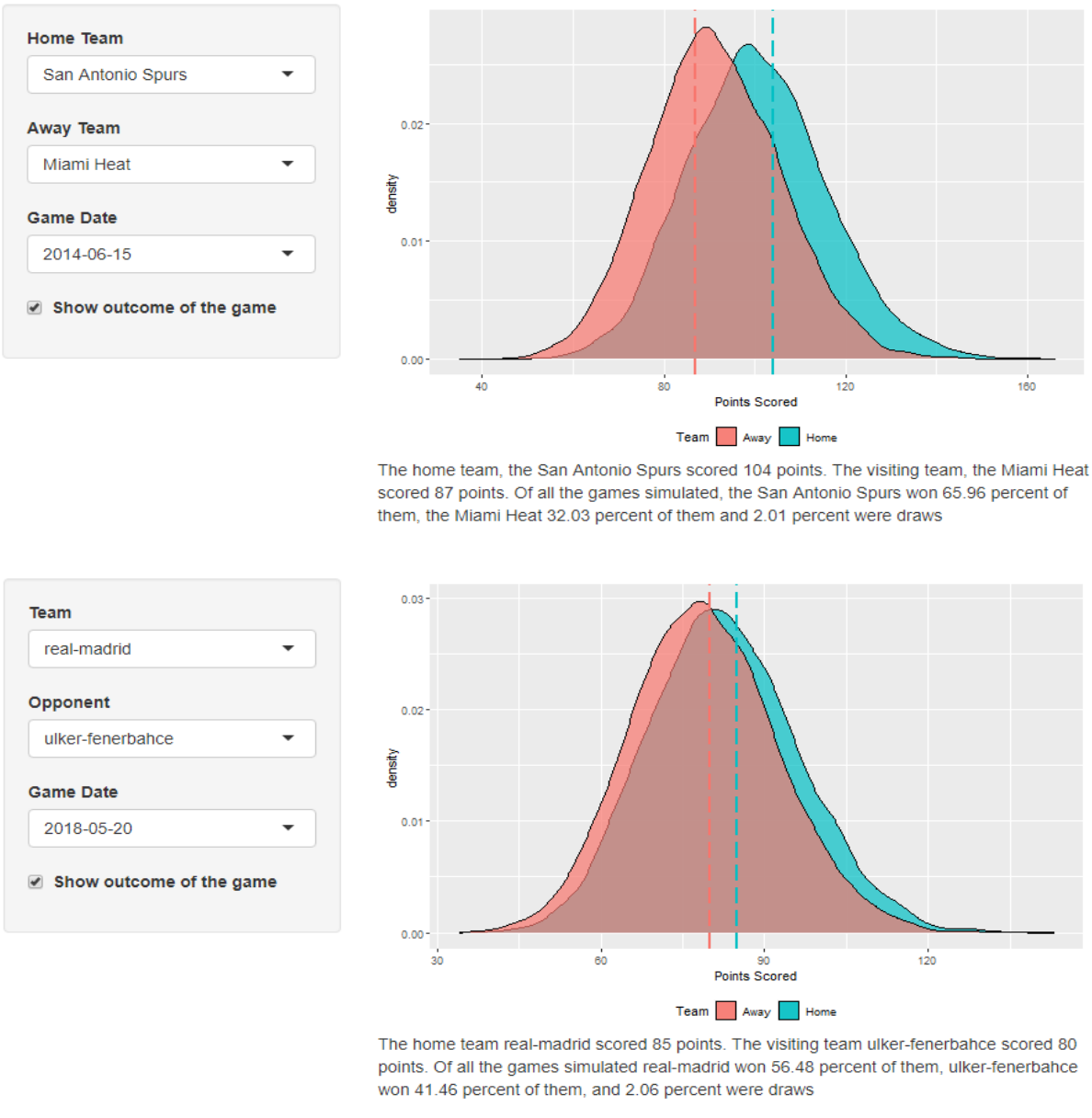


Figure 22. Two examples of the shiny app that display the simulations for a single game. On the top we see the simulations of the last game of the 2014 NBA Finals between the San Antonio Spurs and the Miami Heat on the 14th of June 2014. The bottom shows the outcomes of the simulations of the 2018 EuroLeague finals between Real Madrid and Fenerbahce on the 20th of May 2018.

5 Discussion

5.1 Data Exploration

The Charlotte Hornets had a two-season break, from 2002-2004, due to their relocation to New Orleans, creating the New Orleans Hornets franchise. The New Orleans Hornets started their franchise thus in 2002, while a new franchise was created in Charlotte, the Charlotte Bobcats, which started playing in 2004. The New Orleans Hornets rebranded themselves to the New Orleans Pelicans in 2013, so that the Charlotte Bobcats could officially rename themselves the Charlotte Hornets once again in 2014. In the 2011-2012 season there was a lockout, starting from July 1, 2011 until December 8, 2011, effectively cancelling all preseason games and the first six weeks of the regular season [28]. In the 2012-2013 season, the game between the Boston Celtics and the Indiana Pacers on April the 16th was cancelled in the aftermath of the Boston Marathon bombing. The game was not rescheduled because it would not have had any impact on their playoff seedings [29].

Currently, 11 out of the 18 EuroLeague places are held by licensed clubs that have long-term licenses with EuroLeague Basketball and are members of the Shareholders Executive Board. These licensed clubs are currently: Anadolu Efes (Turkey), Baskonia (Spain), CSKA Moscow (Russia), FC Barcelona (Spain), Fenerbahçe (Turkey), Maccabi Tel Aviv (Israel), Olimpia Milano (Italy), Olympiacos (Greece), Panathinaikos (Greece), Real Madrid (Spain), and Žalgiris (Lithuania). These eleven clubs are eleven of the twelve clubs with the most games played in the EuroLeague, with only Unicaja Malaga having more games than Olimpia Milano. Unicaja Malaga is a former holder of a long-term license (A Licence) but lost it in 2015. The EuroLeague introduced a limitation of three A Licences per country going from the season 2015-2016 forward. Unicaja Malaga ended up being the lower ranked team of the four Spanish teams that season, which resulted in the loss of their license. The remaining seven EuroLeague places are held by five associated clubs that have annual licences and two two-year wild cards. The five associated clubs are awarded through one place going to the winner of the previous season's 2nd-tier European competition, the Euro Cup, with the other four places going to a combination of European national domestic league winners (ABA League, VTB United League, Basketball Bundesliga, Liga ACB). This is the latest format of the EuroLeague.

Overall, the format of the EuroLeague changed quite a bit over the years[4], which was also seen in the data.

Looking at the trends in the different types of play, we find the important and logical observation that in both leagues the number of total plays, zero-point plays, and two-point plays increase when overtimes are played. When implementing the number of overtimes as a variable, the accuracy of the prediction models increased significantly, meaning the number of overtimes explain a significant amount of unexplained variation. The number of overtimes teams are going to play in a game is of course not known beforehand and may be very difficult to predict but doing so may significantly improve the prediction models of these plays. However, a large amount of unexplained variation is still left, even when including overtimes, and thus potentially important variables still need to be found to improve the prediction models.

In general, there has been a steady increase in pace, the average number of plays per game, in the NBA the last decade, which was always higher than in the EuroLeague. This increase in pace was also found in [30]. It comes from the so-called three-point revolution the NBA undergoes, with teams realizing the potential of three-point shots and gradually taking more advantage out of it. Three-point shooting effectiveness is generally lower than two-point shooting effectiveness, which would explain the rise in zero-point plays in the form of missed shots. Attempting more three-point shots has other consequences next to increasing the pace of the game, such as a decrease in two-point field goal attempts. Plays that were previously two-point attempts are now replaced by three-point attempts.

Furthermore, this would also result in less fouls as fouls are more often made within the three-point line. Less fouls means less free-throw attempts resulting in a decrease in one-point plays. Another explanation for the decrease in one-point plays found in both leagues would be the increase in accuracy in free-throw shooting. In recent years, the free-throw accuracy rose to an average of 75% in each league. This means that a personal foul resulting in free-throw attempts results in 1.5 points per possession on average, which is higher than the 1 point per possession for two-point attempts and 1.1 points per possession for three-point attempts [30]. Preventing fouls that result in free-throw attempts would be thus be a sound tactical decision.

The increasing trend in the number of three-point shots is also present, although less dramatically, in the EuroLeague, and also found by [30] and [31].

4 https://en.wikipedia.org/wiki/EuroLeague_historical_league_formats

In the season 2010-2011 however, we saw a major decline in the number of three-point plays compared to the previous seasons, combined with a bump of zero- and two-point plays. This coincides with the major rule changes in the EuroLeague, and several other European and most national basketball federations [31]. These rule changes included, among others, moving the three-point arc 6.75 meters away from the basket, adding a no-charge semicircle under the basket, and changes in how and when the 24-second shot-clock reset. The three-point arc was previously 6.25 meters away from the basket and thus moved back 50 cm. The adjustment period needed for players to adapt to this new distance would explain the decline in three-point plays.

Interesting to see is that the EuroLeague consistently had on average more three-point plays than the NBA, except in the last 2 seasons, while this was never the case in the other type of plays. They even shot and scored on average more three-point shots during the dip after the rule change in season 2010-2011. One reason could be that scoring three-point shots is easier in the EuroLeague compared to the NBA because the distance of the three-point arc from the basket was shorter in the EuroLeague, 6.25 meters before the rule change and 6.75 meters after the rule change in the EuroLeague, compared to 7.24 meters (6.71 meters at the corners) in the NBA. There was however no significant difference in shooting percentage between both leagues according to [30]. The strong increase in NBA three-point plays also seems to counteract that argument. Another argument brought up by [32] is that the NBA is a copycat league and tends to be risk averse.

The importance of three-point shots has been known for years, if not decades, through basketball analytics. The EuroLeague could be considered a league that puts more emphasis on the tactical aspects of basketball, while the NBA puts a lot of value in athleticism. The three-point field goal percentage and three-point field goals made have been a deciding factor for winning a game for many years in the EuroLeague [33][34]. The EuroLeague may therefore have been quicker to adopt a new strategy that gives them an edge over their opponent, while the NBA teams would have wanted to see the evidence of the tactic in the winning teams first before adopting it [32]. The success of the record-breaking Golden State Warriors since the season 2014-2015, under career three-point field goal percentage record holder Steve Kerr as rookie head coach and with three-point shooting machines Stephen Curry and Klay Thompson as players, was then likely the ultimate catalyst that changed the NBA into the three-point league that it is today.

5.2 Predicting Plays

The models trying to predict the different types of plays performed poorly. This was not unexpected due the number of possible factors that can influence them. Many of these factors are easy to imagine but impossible to quantify or know beforehand. One factor that was already discussed above is overtime, which unsurprisingly proved to be a huge influence. Another example of such a factor is team tactics. These tactics are determined before the start of the game by the coach but will be changed on the fly during the game if necessary. A team that is expected to lose because they are just not the better team will often try to find a risky tactic that still gives them a shot to win. This usually involves dramatically changing their play style which will result in a very different number of plays than expected. The list goes on. Nevertheless, even poorly performing models gave an improvement in the simulations predicting the game outcomes. Improving these models by finding relevant new variables, using different or using more advanced modelling techniques to predict the types of plays could improve the already good game outcome predictions.

The outcome of these models still gave interesting insights about the factors influencing the different types of plays in both leagues, and differences between the leagues. The most important variables were more easily identified for the NBA than for the EuroLeague due to the availability of more data and the fact that the same teams played in the NBA each season. If a variable is part of the model for all 30 NBA teams, its influence compared to a variable only present in the model of 14 NBA teams is clear. In the EuroLeague, only around a fifth of the teams have played enough games the last ten seasons for a useable prediction model. We thus found that the most common variables are shared between at most 15 teams and sometimes as little as five teams.

Home advantage plays a big role in both leagues. It was the most important factor reducing the number of zero-point plays, while it was the most important factor improving the number of two-point plays. Home advantage also had a positive influence for many teams on the total plays, one-point plays, and three-point plays. These findings are in line with [35], who found that the key factor in home advantage is style of play. Teams that take more two-point and free-throw shots see larger home advantages in the NBA. They argue that the rise in three-point shooting in recent year could partially explain the gradual decline in home advantage. Similar results were found in the EuroLeague by [36]. They found that home teams are characterised by a higher number of assists, steals and points, while road teams have more turnovers.

A difference with the NBA is that wherever there was an influence of home advantage, there was an opposite influence of consecutive road games the team plays in the EuroLeague. This reinforces the influence of home advantage in the EuroLeague and potentially making it stronger than in the NBA. Road games, and especially being on the road a longer time for consecutive road games, has an impact on the performance of teams. Travelling is tiring, which is reinforced by crossing time zones [37][38]. NBA franchises have bigger budgets than EuroLeague teams, with the lowest value NBA franchise worth 1.3 billion dollars [39], while the highest value EuroLeague team is only worth 41 million euros [40], only slightly more than the salary of the highest paid NBA player, Stephen Curry, this season 2019-2020[5]. This difference in budget could mean a difference in travel comfort between both leagues. There is also a difference in arenas. The NBA arenas are all very similar, both in capacity and built year[6], while EuroLeague arenas are very different from each other, both in capacity and built year[7]. All this could potentially help explain the differences in home advantage between the NBA and EuroLeague.

Another difference between both leagues is the factors influencing zero-point plays. In the NBA, the average number of steals of the opponent is the most important, followed closely by defensive rebounds and blocks, basically all the most important defensive statistics. The importance of steals is accentuated by its negative impact on the number of two-point plays. In the EuroLeague, the average number of blocks of the opponent is the most important, with defensive rebounds and steals much further down the list.

In general, it seems that variables related to the opponent the team is playing against are more important than team variables in predicting the different types of plays. In other words, the opponent team will mostly determine if the team will get their average in the different types of plays or not. Playing against fast-paced opponents will result in more plays for the team overall. Playing against good rebounding teams will result in less plays. Playing against good defensive teams will result in more zero-point plays and less two- and three-point plays. Playing against aggressive teams committing a lot of personal fouls will result in more one-point plays. The exception however lays in the average number of three-point plays. A team simply attempting more three-point shots and/or having a higher three-point shot rate in a season than they do on average across seasons will result in more three-point plays in their games that season.

5 https://www.basketball-reference.com/contracts/
6 https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_arenas
7 https://en.wikipedia.org/wiki/List_of_EuroLeague_arenas

5.3 Simulations

Our simulations worked very well in predicting outcomes of both NBA and EuroLeague games. The prediction accuracy of our NBA training set, 65.4%, and our NBA test season 2018-2019, 67.3%, is similar to or even higher than the achieved accuracies by several papers using various machine learning methods [2][3][7][11][41][42][43][44]. The betting market and the experts are still slightly more accurate than our simulation model (68.7% – 69.4%) [7][11][42]. One paper also showcased the power of neural networks in predicting NBA games by achieving an accuracy of 74%, much higher than any other prediction accuracies [7]. The accuracy of EuroLeague games is generally higher, which was also the case for us with 71.1% accuracy for our training set, and 78.6% accuracy for our 2018-2019 test season. This is better than most papers attempting to predict EuroLeague game outcomes. An author compared several different machine learning methods to predict the outcome of EuroLeague games and did not exceed an accuracy of 67%. Additionally, he demonstrated the "wisdom of the basketball crowd", where he asked members of a basketball forum to predict the winner of games, which was decided by majority vote. This resulted in an accuracy of 73%, higher than the machine learning models he used [45]. Another paper used a Poisson regression to predict the scores of EuroLeague games and obtained a game-winning team accuracy of 73% in their test season 2011-2012 [13]. The last paper used a k-nearest neighbours classification to predict the outcome of EuroLeague games and obtained and accuracy of 83.96%. This prediction accuracy is surprisingly high [1].

Our simulations based on the team's seasonal average number of plays only performed slightly worse, give or take 3%, than the simulations based on the predicted values of the number of plays, in both leagues. This is not surprising given the poor fit of the prediction models, but it does make this approach so interesting. There is no need for advanced statistical modelling, and it saves a lot of time and computing power while getting similar results. If you have the data, which is freely available on the internet in the form of basic play-by-play data, simulating a game takes only seconds and its interpretation is very simple. The team with less zero-point plays, and more one-, two-, three-point plays wins more on average. The simulations are also easily visualisable with two normal distribution curves representing the probability density function of the amount of points scored by each opposing team. This can serve as an extra tool when trying to predict single games. Finally, the fact that our game prediction accuracy is higher when using the predicted plays from our poorly performing prediction models indicates that improving these prediction models can potentially increase the accuracy even more.

6 Conclusion

The prediction models attempting to predict the different types of plays did not perform very well, giving only slightly better predictions than the average. Much of the variation was left unexplained, with the main identified excluded variable for the total plays, zero-point plays and two-point plays being the number of overtimes. It still provided some insights in the factors affecting the different kinds of plays and their differences between the NBA and the EuroLeague. Home advantage plays a big role in both leagues, especially for the zero-point plays and the two-point plays. The impact of consecutive road games reinforced the influence of home advantage in the EuroLeague while it did less so in the NBA. In general, the statistics of the opponent had more impact on the different types of plays than the statistics of the team itself, except for the three-point plays.

The simulations performed very well in predicting the game winners in both NBA and EuroLeague. The prediction accuracy for the NBA was comparable, if not better, than most prediction models out there, but still performed slightly worse than the experts and the betting market. The prediction accuracy for the EuroLeague was higher than other prediction models found. The simulations based on the non-predicted play data were only 3% less accurate than the simulations based on the predicted plays. This shows the strength of that method, as it can give very good predictions without resorting to complex machine learning models. It is fast, easy to use, interpret and visualize. The increase of 3% in accuracy using the play data from our poorly performing prediction models shows that there is room for improvement. If the prediction models can be optimized, the prediction accuracy is expected to become even better than it already is. Compound Poisson simulation is thus a powerful and accessible way of predicting outcomes of basketball games and worth further research.

7 References

Videos

a. https://www.youtube.com/watch?v=eYwjjo3cC70 (from 1:53 on, watched 30/07/2020)
b. https://www.youtube.com/watch?v=uLD8pqMPPQo (watched 30/07/2020)
c. https://www.youtube.com/watch?v=qv7GbA0R76U (watched 30/07/2020)

Papers

1. Horvat, T., Job, J., & Medved, V. (2018, January). Prediction of Euroleague games based on supervised classification algorithm k-nearest neighbours. In *6th International Congress on Sport Sciences Research and Technology Support*.
2. Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010, September). The use of data mining for basketball matches outcomes prediction. In *IEEE 8th International Symposium on Intelligent Systems and Informatics* (pp. 309-312). IEEE.
3. Cao, C. (2012). Sports data mining technology used in basketball outcome prediction.
4. Caudill, S. B. (2003). Predicting discrete outcomes with the maximum score estimator: The case of the NCAA men's basketball tournament. *International Journal of Forecasting*, *19*(2), 313-317.
5. Shi, Z., Moorthy, S., & Zimmermann, A. (2013, September). Predicting NCAAB match outcomes using ML techniques–some results and lessons learned. In *ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics*.
6. Beckler, M., Wang, H., & Papamichael, M. (2013). Nba oracle. *Zuletzt besucht am*, *17*(20082009.9).
7. Loeffelholz, B., Bednar, E., & Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports*, *5*(1).
8. Vračar, P., Štrumbelj, E., & Kononenko, I. (2016). Modeling basketball play-by-play data. *Expert Systems with Applications*, *44*, 58-66.
9. Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, *28*(2), 532-542.

10. Shirley, K. (2007, September). A Markov model for basketball. In *New England Symposium for Statistics in Sports* (pp. 82-82).
11. Peuter, C. D. (2013). *Modeling Basketball Games as Alternating Renewal-Reward Processes and Predicting Match Outcomes* (Doctoral dissertation, Master's thesis, Duke University).
12. Gabel, A., & Redner, S. (2012). Random walk picture of basketball scoring. *Journal of Quantitative Analysis in Sports*, *8*(1).
13. Alp Erilli, N., Ermis, E., & Yalcin Tasmektepligil, M. (2013). Basketball "Turkish Airlines EuroLEague" 2011-12 season Poisson regression simulation modelling.. *International Journal Of Academic Research*, *5*(5).
14. Oliver, D. (2004). *Basketball on paper: rules and tools for performance analysis*. Potomac Books, Inc..
15. Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, *3*(3).
16. Sergio O. P. (2019). Eurolig. GitHub repository, https://github.com/solmos/eurolig .
17. Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, *3*(3).
18. Oliver, D. (2004). *Basketball on paper: rules and tools for performance analysis*. Potomac Books, Inc..
19. Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment*, *91*(2), 121-136.
20. Liu, E. (2019). Deviance of Poisson regression.
21. Cameron, A. C., & Windmeijer, F. A. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, *14*(2), 209-220.
22. Hastie, T., & Qian, J. (2016). An Introduction to glmnet.
23. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.
24. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.

25. Mwikali, J., Mwalili, S., & Wanjoya, A. (2019). Penalized Poisson Regression Model Using Elastic Net and Least Absolute Shrinkage and Selection Operator (Lasso) Penality. *International Journal of Data Science and Analysis*, *5*(5), 99.
26. Agresti, A. (2003). *Categorical data analysis* (Vol. 482). John Wiley & Sons.
27. Zhang, Y. W. (2018). CPLM: compound Poisson linear models.
28. Staudohar, P. D. (2012). The basketball lockout of 2011. *Monthly Lab. Rev.*, *135*, 31.
29. Golliver B. (2013). NBA cancels game between Celtics and Pacers after Boston Marathon blasts. *Sports Illustrated (Website).* Retrieved on 17/08/2020, from https://www.si.com/nba/2013/04/16/boston-marathon-bombing-terror-attack-celtics-pacers-game-cancelled-nba
30. Mandić, R., Jakovljević, S., Erčulj, F., & Štrumbelj, E. (2019). Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017. *PloS one*, *14*(10), e0223524.
31. Štrumbelj, E., Vračar, P., Robnik-Šikonja, M., Dežman, B., & Erčulj, F. (2013). A decade of euroleague basketball: An analysis of trends and recent rule change effects. *Journal of human kinetics*, *38*(2013), 183-189.
32. Chung, J. Explaining the Trends of NBA Strategy through the Lens of Human Risk Tolerance.
33. Çene, E. (2018). What is the difference between a winning and a losing team: insights from Euroleague basketball. *International Journal of Performance Analysis in Sport*, *18*(1), 55-68.
34. Dogan, I., & Ersoz, Y. (2019). The important game-related statistics for qualifying next rounds in Euroleague. *Montenegrin Journal of Sports Science and Medicine*, *8*(1), 43-50.
35. Harris, A. R., & Roebber, P. J. (2019). NBA team home advantage: Identifying key factors using an artificial neural network. *PloS one*, *14*(7), e0220630.
36. Pojskić, H., Šeparović, V., & Užičanin, E. (2011). Modelling home advantage in basketball at different levels of competition. *Acta Kinesiologica*, *5*(1), 25-30.
37. Swartz, T. B., & Arce, A. (2014). New insights involving the home team advantage. *International Journal of Sports Science & Coaching*, *9*(4), 681-692.
38. Entine, O. A., & Small, D. S. (2008). The role of rest in the NBA home-court advantage. *Journal of Quantitative Analysis in Sports*, *4*(2).
39. Badenhausen K., Ozanian M., & Settimi C. (2020). NBA team values 2020: Lakers and Warriors join Knicks un rarefield $4 billion club. *Forbes (Website).* Retrieved on

17/08/2020, from https://www.forbes.com/sites/kurtbadenhausen/2020/02/11/nba-team-values-2020-lakers-and-warriors-join-knicks-in-rarefied-4-billion-club/

40. Askounis J. (2019). List of season budgets of EuroLeague teams according to L'Equipe. *Eurohoops (Website)*. Retrieved on 17/08/2020, from https://www.eurohoops.net/en/euroleague/963346/list-of-season-budgets-of-euroleague-teams/
41. Zimmermann, A. (2016). Basketball predictions in the NCAAB and NBA: Similarities and differences. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(5), 350-364.
42. Manner, H. (2016). Modeling and forecasting the outcomes of NBA basketball games. *Journal of Quantitative Analysis in Sports*, *12*(1), 31-41.
43. Horvat, T., & Job, J. (2019). Importance of the training dataset length in basketball game outcome prediction by using naive classification machine learning methods. *Elektrotehniski Vestnik*, *86*(4), 197-202.
44. Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *International Journal of Forecasting*, *28*(2), 532-542.
45. Giasemidis, G. (2020). Descriptive and Predictive Analysis of Euroleague Basketball Games and the Wisdom of Basketball Crowds. *arXiv preprint arXiv:2002.08465*.

8 Appendix

Appendix 1

NBA Abbreviations

Team Names	Team Abbreviations
Atlanta Hawks	ATL
Boston Celtics	BOS
Brooklyn Nets	NJN
Charlotte Hornets	CHA
Chicago Bulls	CHI
Cleveland Cavaliers	CLE
Dallas Mavericks	DAL
Denver Nuggets	DEN
Detroit Pistons	DET
Golden State Warriors	GSW
Houston Rockets	HOU
Indiana Pacers	IND
Los Angeles Clippers	LAC
Los Angeles Lakers	LAL
Memphis Grizzlies	MEM
Miami Heat	MIA
Milwaukee Bucks	MIL
Minnesota Timberwolves	MIN
New Orleans Pelicans	NOH/NOP
New York Knicks	NYK
Oklahoma City Thunder	OKC
Orlando Magic	ORL
Philadelphia 76ers	PHI
Phoenix Suns	PHO
Portland Trail Blazers	POR
Sacremento Kings	SAC
San Antonio Spurs	SAS
Toronto Raptors	TOR
Utah Jazz	UTA
Washington Wizards	WAS

EuroLeague Abbreviations

Team Names	Team Abbreviations
AEK Athens	aek-athens
ALBA Berlin	alba-berlin
Alba Berlin	alba-berlin
Anadolu Efes	anadolu-efes
Anadolu Efes Istanbul	anadolu-efes
Efes Pilsen	anadolu-efes
Efes Pilsen Istanbul	anadolu-efes
Aris TT Bank	aris
Air Avellino	avellino
AXA FC Barcelona	barcelona
FC Barcelona	barcelona
FC Barcelona Lassa	barcelona
FC Barcelona Regal	barcelona
Regal Barcelona	barcelona
Regal FC Barcelona	barcelona
Winterthur FC Barcelona	barcelona
Bayern Munich	bayern-muenchen
FC Bayern Munich	bayern-muenchen
Besiktas Integral Forex	besiktas
Besiktas JK Istanbul	besiktas
Gescrap Bilbao Basket	bilbao
Gescrap Bizkaia Bilbao Basket	bilbao
Brose Bamberg	brose-baskets
Brose Baskets	brose-baskets
Brose Baskets Bamberg	brose-baskets
GHP Bamberg	brose-baskets
BC Budivelnyk	budivelnik
Buducnost	buducnost
Buducnost VOLI	buducnost
Buducnost VOLI Podgorica	buducnost
KK Buducnost	buducnost
Bennet Cantu	cantu
Mapooro Cantu	cantu
Cedevita Zagreb	cedevita

Elan Chalon	chalon
Elan Chalon-Sur-Saone	chalon
Cholet Basket	cholet
Cibona	cibona-zagreb
Cibona VIP	cibona-zagreb
Cibona Zagreb	cibona-zagreb
CSKA Moscow	cska-moscow
Darussafaka Dogus	darussafaka
Darussafaka Dogus Istanbul	darussafaka
Darussafaka Tekfen	darussafaka
Darussafaka Tekfen Istanbul	darussafaka
Dynamo Moscow	dynamo-moscow
Adecco Estudiantes	estudiantes
Estudiantes	estudiantes
Climamio Bologna	fortitudo-bologna
Climamio Fortitudo Bologna	fortitudo-bologna
PAF Bologna	fortitudo-bologna
Paf Wennington Fortitudo Bologna	fortitudo-bologna
Skipper Bologna	fortitudo-bologna
Skipper Fortitudo Bologna	fortitudo-bologna
Opel Skyliners	frankfurt
Galatasaray Liv Hospital	galatasaray
Galatasaray Liv Hospital Istanbul	galatasaray
Galatasaray Medical Park	galatasaray
Galatasaray Odeabank	galatasaray
Galatasaray Odeabank Istanbul	galatasaray
Herbalife Gran Canaria	gran-canaria
Hapoel Jerusalem	hapoel-jerusalem
DKV Joventut	joventut
Pinar Karsiyaka Izmir	karsiyaka
BC Khimki	khimki
Khimki	khimki
Khimki Moscow Region	khimki
RheinEnergie	koln
RheinEnergie Koln	koln
Lokomotiv Kuban Krasnodar	kuban

Lokomotiv-Kuban	kuban
Le Mans	le-mans
Le Mans Sarthe Basket	le-mans
Limoges CSP	limoges
Haribo London Towers	london-towers
Kinder London Towers	london-towers
London Towers	london-towers
Lugano Snakes	lugano
Lugano Tigers	lugano
Maccabi Electra	maccabi-tel-aviv
Maccabi Electra Tel Aviv	maccabi-tel-aviv
Maccabi Elite	maccabi-tel-aviv
Maccabi Elite Tel Aviv	maccabi-tel-aviv
Maccabi FOX Tel Aviv	maccabi-tel-aviv
Maroussi	maroussi
Maroussi B.C.	maroussi
AJ Milano	milano
Armani Jeans Milano	milano
Armani Jeans Milano	milano
AX Armani Exchange Olimpia	milano
AX Armani Exchange Olimpia Milan	milano
EA7 Emporio Armani	milano
EA7 Emporio Armani Milan	milano
EA7-Emporio Armani Milano	milano
SLUC Nancy	nancy
JSF Nanterre	nanterre
Eldo Napoli	napoli
Neptunas Klaipeda	neptunas
Nizhny Novgorod	novgorod
KK Krka	novo-mesto
EWE Baskets	oldenburg
EWE Oldenburg	oldenburg
KK Union Olimpija	olimpija-ljubljana
Olympiacos	olympiakos
Olympiacos Piraeus	olympiakos
Olympiakos	olympiakos

BC Oostende	oostende
Orleans Loiret Basket	orleans
Ovarense	ovarense
Ovarense Aerosoles	ovarense
Panathinaikos	panathinaikos
Panathinaikos Athens	panathinaikos
Panathinaikos OPAP	panathinaikos
Panathinaikos OPAP Athens	panathinaikos
Panathinaikos Superfoods	panathinaikos
Panathinaikos Superfoods Athens	panathinaikos
Panionios On Telecoms	panionios
PAOK Thessaloniki	paok
Partizan	partizan
Partizan Igokea	partizan
Partizan mt:s Belgrade	partizan
Partizan NIS Belgrade	partizan
Elan Bearnais Pau-Orthez	pau-orthez
Peristeri	peristeri
Ural Great	perm
Ural Great Perm	perm
Scavolini Pesaro	pesaro
Asseco Prokom	prokom
Asseco Prokom Gdynia	prokom
Asseco Prokom Sopot	prokom
Prokom Trefl Sopot	prokom
Real Madrid	real-madrid
Crvena zvezda / Red Star	red-star
Crvena Zvezda mts	red-star
Crvena Zvezda mts Belgrade	red-star
Crvena Zvezda Telekom Belgrade	red-star
Chorale Roanne	roanne
Lottomatica Roma	roma
Banco di Sardegna Sassari	sassari
Dinamo Banco di Sardegna Sassari	sassari
Montepaschi Siena	siena
Belgacom Spirou Charleroi	spirou

Region Wallone Spirou Charleroi	spirou
Spirou Basket	spirou
Spirou Charleroi	spirou
Saint Petersburg Lions	st-petersburg
Strasbourg	strasbourg
Strasbourg IG	strasbourg
Benetton Basket	treviso
Benetton Treviso	treviso
PGE Turow Zgorzelec	turow
Fenerbahce	ulker-fenerbahce
Fenerbahce Beko Istanbul	ulker-fenerbahce
Fenerbahce Dogus Istanbul	ulker-fenerbahce
Fenerbahce Istanbul	ulker-fenerbahce
Fenerbahce Ulker	ulker-fenerbahce
Fenerbahce Ulker Istanbul	ulker-fenerbahce
Ulker	ulker-fenerbahce
Unicaja	unicaja-malaga
Unicaja Malaga	unicaja-malaga
UNICS	unics-kazan
UNICS Kazan	unics-kazan
Pamesa Valencia	valencia
Power Electronics Valencia	valencia
Valencia Basket	valencia
Muller Verona	verona
Asvel Basket	villeurbanne
ASVEL Villeurbanne	villeurbanne
Adecco ASVEL Villeurbanne	villeurbanne
Lietuvos rytas	vilnius
Lietuvos Rytas	vilnius
Lietuvos Rytas Vilnius	vilnius
Kinder Bologna	virtus-bologna
VidiVici Bologna	virtus-bologna
Virtus Bologna	virtus-bologna
Virtus VidiVici	virtus-bologna
Baskonia	vitoria
Baskonia Vitoria Gasteiz	vitoria

Caja Laboral	vitoria
Caja Laboral Vitoria	vitoria
Kirolbet Baskonia	vitoria
Laboral Kutxa	vitoria
Laboral Kutxa Vitoria	vitoria
Laboral Kutxa Vitoria Gasteiz	vitoria
TAU Ceramica	vitoria
Tau Ceramica	vitoria
Slask Wroclaw	wroclaw
KK Zadar	zadar
KK Zagreb	zagreb
BC Zalgiris	zalgiris
Zalgiris	zalgiris
Zalgiris Kaunas	zalgiris
Zalgiris Kaunus	zalgiris
Stelmet Zielona Gora	zielona-gora
Budivelnik Kiev	budivelnik
Entente Orleanaise	orleans
Idea Slask	wroclaw
KIROLBET Baskonia Vitoria Gasteiz	vitoria
KIROLBET Baskonia Vitoria-Gasteiz	vitoria
KRKA Novo Mesto	novo-mesto
Pau-Orthez	pau-orthez
Telindus Oostende	oostende
Union Olimpija	olimpija-ljubljana

Appendix 2

Basic Box Scores

FG	Field Goals Made
FGA	Field Goals Attempted
FG%	Field Goals Percentage
3P	3-Point Field Goals Made
3PA	3-Point Field Goals Attempted
3P%	3-Point Field Goals Percentage
2P	2-Point Field Goals Made
2PA	2-Point Field Goals Attempted
2P%	2-Point Field Goals Percentage
FT	Free Throws Made
FTA	Free Throws Attempted
FT%	Free Throws Percentage
ORB	Offensive Rebounds
DRB	Defensive Rebounds
TRB	Total Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal Fouls
PTS	Points Made

Advanced Boxscores

TS%	True Shooting Percentage: A measure of shooting efficiency that takes into account 2-point field goals, 3-point fields goals, and free throws
3PAR	3-Point Attempt Rate: Percentage of FG Attempts from 3-Point Range
FTr	Free Throw Attempt Rate: Number of FT Attempts Per FG Attempt
ORB%	Offensive Rebound Percentage: An estimate of the percentage of available offensive rebounds a team grabbed
DRB%	An estimate of the percentage of available defensive rebounds a team grabbed
TRB%	Total Rebound Percentage: An estimate of the percentage of available rebounds a team grabbed
AST%	Assist Percentage: An estimate of the percentage of teammate field goals a team assisted
STL%	Steal Percentage: An estimate of the percentage of opponent possessions that end with a steal by the team
BLK%	Block Percentage: An estimate of the percentage of opponent two-point field goal attempts blocked by the team
TOV%	Turnover Percentage: An estimate of turnovers committed per 100 plays
ORtg	Offensive Rating: An estimate of points produced or scored per 100 possessions
DRtg	Defensive Rating: An estimate of points allowed per 100 possessions
eFG%	Effective Field Goal Percentage: This statistic adjust for the fact that 3-point field goal is worth one more point than a 2-point goal

Appendix 3

Game IDs of inconsistent EuroLeague games:

2008-01-31_ cibona-zagreb, 2008-03-19_partizan, 2008-11-27_ olimpija-ljubljana, 2008-12-11_real-madrid, 2009-10-29_orleans, 2010-03-30_real-madrid, 2010-12-01_khimki, 2011-02-17_olimpija-ljubljana, 2012-03-23_siena, 2012-11-09_olimpija-ljubljana, 2015-11-06_unicaja-malaga, 2017-10-19_milano, 2018-02-02_brose-baskets, 2018-11-15_darussafaka, 2018-11-15_real-madrid, 2019-04-24_barcelona

Game IDs of inconsistent NBA games:

200111070TOR, 200204070PHI, 200204160ATL, 200211050CLE, 200212130BOS, 200302220MIA, 200402040WAS, 200402180NOH, 200502270PHO, 200504030CLE, 200504170MIA, 200505070BOS, 200511190DAL, 200811180GSW, 200902170PHO, 200911220MIA,201103310LAL, 201201270BOS, 201202170HOU, 201212250MIA, 201302030DET, 201404020NYK, 201503160GSW, 201504120DEN, 201512060DET, 201612190DEN, 201701200LAL, 201701250BOS, 201710210UTA, 201710220BRK, 201711100PHO, 201711250GSW, 201711290PHI, 201712060BOS, 201712110HOU, 201801040LAC, 201801100CHO, 201801150BRK, 201801220LAC, 201802060GSW, 201802070MEM, 201803150POR, 201803170NOP, 201901050DET

Appendix 4

Variables used in the models

game_id	Unique ID of the game
Season	Year in which the game's season finishes
GameType	Regular season game or playoff game
Team	Abbreviation of the team
Team_FG	Average number of field goals made by the team this season
Team_FGA	Average number of field goals attempted by the team this season
Team_FG%	Average percentage of field goals made by the team this season
Team_2P	Average number of two-point field goals made by the team this season
Team_2PA	Average number of two-point field goals attempted by the team this season
Team_2P%	Average percentage of two-point field goals made by the team this season
Team_3P	Average number of three-point field goals made by the team this season
Team_3PA	Average number of three-point field goals attempted by the team this season
Team_3P%	Average percentage of three-point field goals made by the team this season
Team_FT	Average number of free throws made by the team this season
Team_FTA	Average number of free throws attempted by the team this season
Team_FT%	Average percentage of free throws made by the team this season
Team_ORB	Average number of offensive rebounds by the team this season
Team_DRB	Average number of defensive rebounds by the team this season
Team_TRB	Average number of total rebounds by the team this season

Team_AST	Average number of assists by the team this season
Team_STL	Average number of steals by the team this season
Team_BLK	Average number of blocks by the team this season
Team_TOV	Average number of turnovers by the team this season
Team_PF	Average number of personal fouls by the team this season
Team_PTS	Average number of points made by the team this season
Team_TS%	Average true shooting percentage of the team this season
Team_eFG%	Average effective field goal percentage of the team this season
Team_3Par	Average 3-point attempt rate of the team this season
Team_FTr	Average free-throw attempt rate of the team this season
Team_ORB%	Average offensive rebound percentage of the team this season
Team_DRB%	Average defensive rebound percentage of the team this season
Team_TRB%	Average total rebound percentage of the team this season
Team_AST%	Average assist percentage of the team this season
Team_BLK%	Average block percentage of the team this season
Team_ORtg	Team Offensive Rating
Team_DRtg	Team Defensive Rating
Opponent	Abbreviation of the opponent team
Opp_FG	Average number of field goals made by the opponent this season
Opp_FGA	Average number of field goals attempted by the opponent this season
Opp_FG%	Average percentage of field goals made by the opponent this season
Opp_2P	Average number of two-point field goals made by the opponent this season
Opp_2PA	Average number of two-point field goals attempted by the opponent this season

Opp_2P%	Average percentage of two-point field goals made by the opponent this season
Opp_3P	Average number of three-point field goals made by the opponent this season
Opp_3PA	Average number of three-point field goals attempted by the opponent this season
Opp_3P%	Average percentage of three-point field goals made by the opponent this season
Opp_FT	Average number of free throws made by the opponent this season
Opp_FTA	Average number of free throws attempted by the opponent this season
Opp_FT%	Average percentage of free throws made by the opponent this season
Opp_ORB	Average number of offensive rebounds by the opponent this season
Opp_DRB	Average number of defensive rebounds by the opponent this season
Opp_TRB	Average number of total rebounds by the opponent this season
Opp_AST	Average number of assists by the opponent this season
Opp_STL	Average number of steals by the opponent this season
Opp_BLK	Average number of blocks by the opponent this season
Opp_TOV	Average number of turnovers by the opponent this season
Opp_PF	Average number of personal fouls by the opponent this season
Opp_PTS	Average number of points made by the opponent this season
Opp_TS%	Average true shooting percentage of the opponent this season
Opp_eFG%	Average effective field goal percentage of the opponent this season
Opp_3Par	Average 3-point attempt rate of the opponent this season
Opp_FTr	Average free-throw attempt rate of the opponent this season
Opp_AST%	Average assist percentage of the opponent this season
Opp_BLK%	Average block percentage of the opponent this season

Team_0PtsPoss	Average number of zero point plays by the team this season
Team_1PtsPoss	Average number of one point plays by the team this season
Team_2PtsPoss	Average number of two point plays by the team this season
Team_3PtsPoss	Average number of three point plays by the team this season
Team_4PtsPoss	Average number of four point plays by the team this season
Team_TotalPoss	Average number of total plays by the team this season
Team_TeamPoss	Average number of team possessions by the team this season
Opp_0PtsPoss	Average number of zero point plays by the opponent this season
Opp_1PtsPoss	Average number of one point plays by the opponent this season
Opp_2PtsPoss	Average number of two point plays by the opponent this season
Opp_3PtsPoss	Average number of three point plays by the opponent this season
Opp_4PtsPoss	Average number of four point plays by the opponent this season
Opp_TotalPoss	Average number of total plays by the opponent this season
Homeadvantage	Team plays at home (TRUE) or does not play at home (FALSE)
Team_Wins	Number of games the team won this season
Team_Losses	Number of games the team lost this season
Opp_Wins	Number of games the opponent won this season
Opp_Losses	Number of games the opponent lost this season

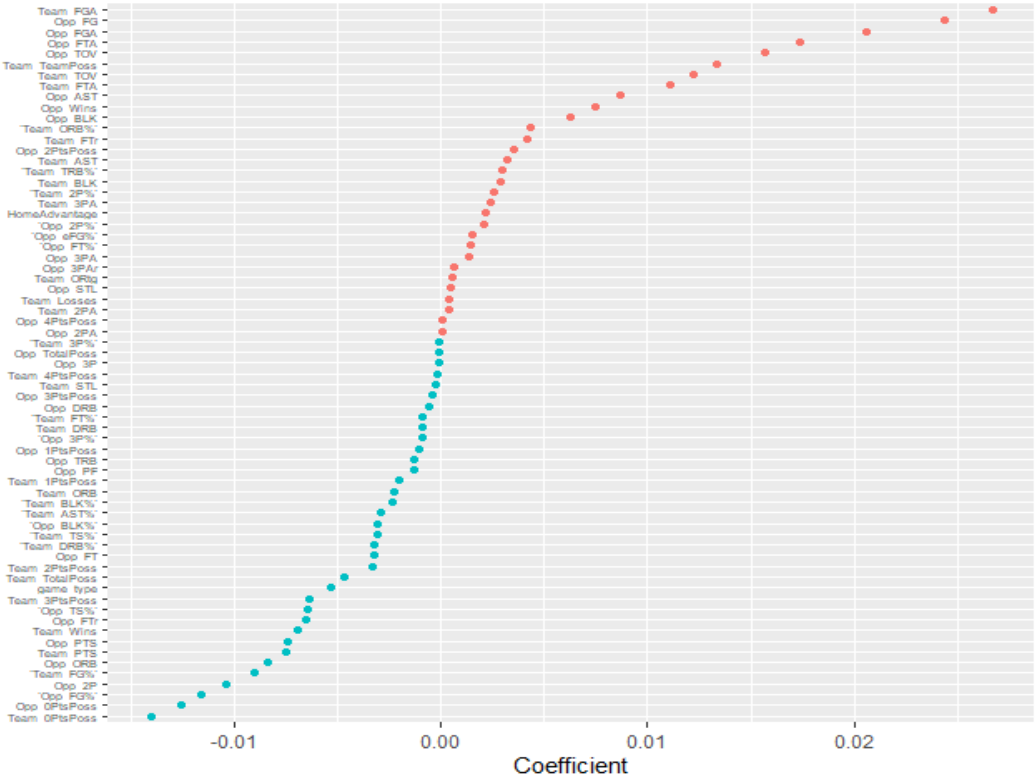
Note the abbreviation Poss in some of the variables. The variables were created when I still used the term possessions instead of plays but they indicate plays, except for the variable Team_TeamPoss which refers to the team possessions as defined in the methods. Still need to be changed.

Appendix 5

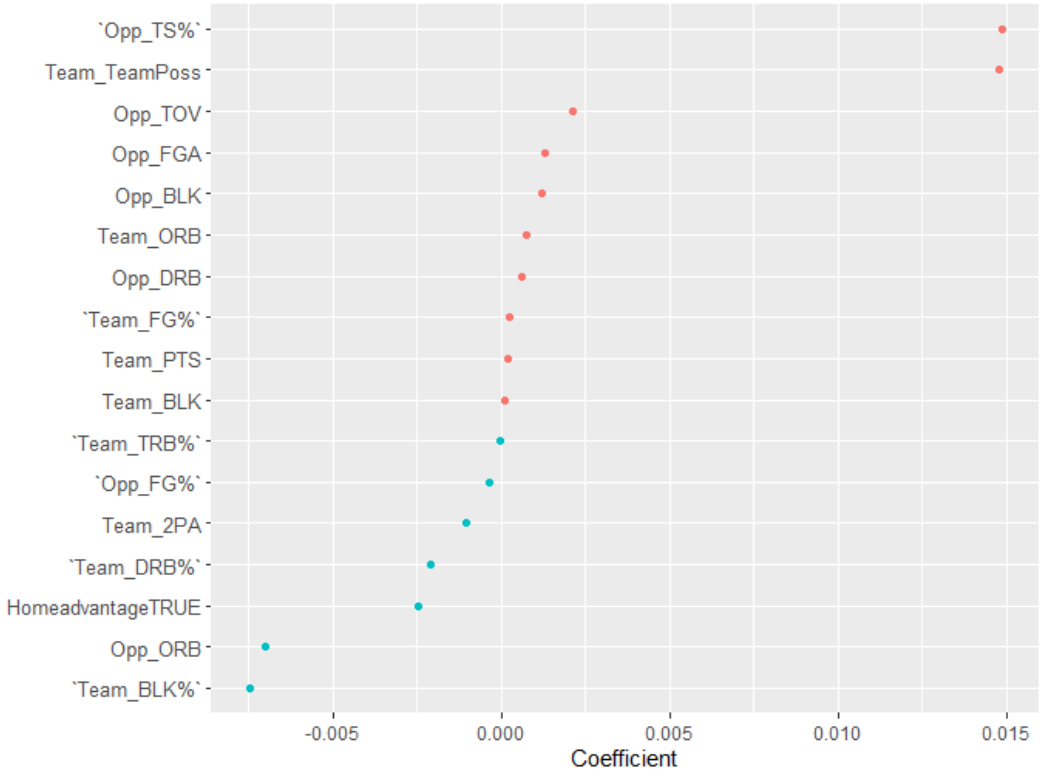
The significant variables in the general models. The categorical variables Season, Team and Opponent are not shown in the figures to save space. The red points show positive coefficients, while the blue points show the negative coefficients. The true number of variables, including the previously mentioned season and team variables are mentioned between brackets.

Total Plays

NBA (150 variables)

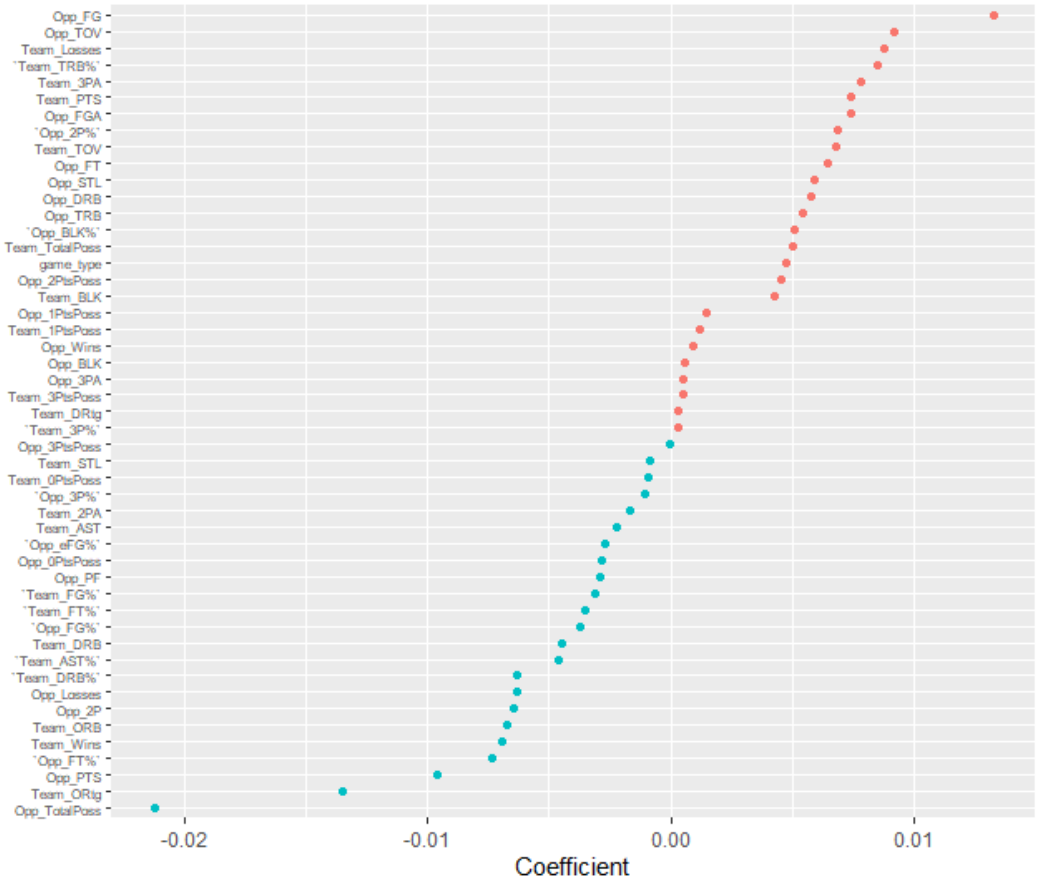


EuroLeague (54 variables)

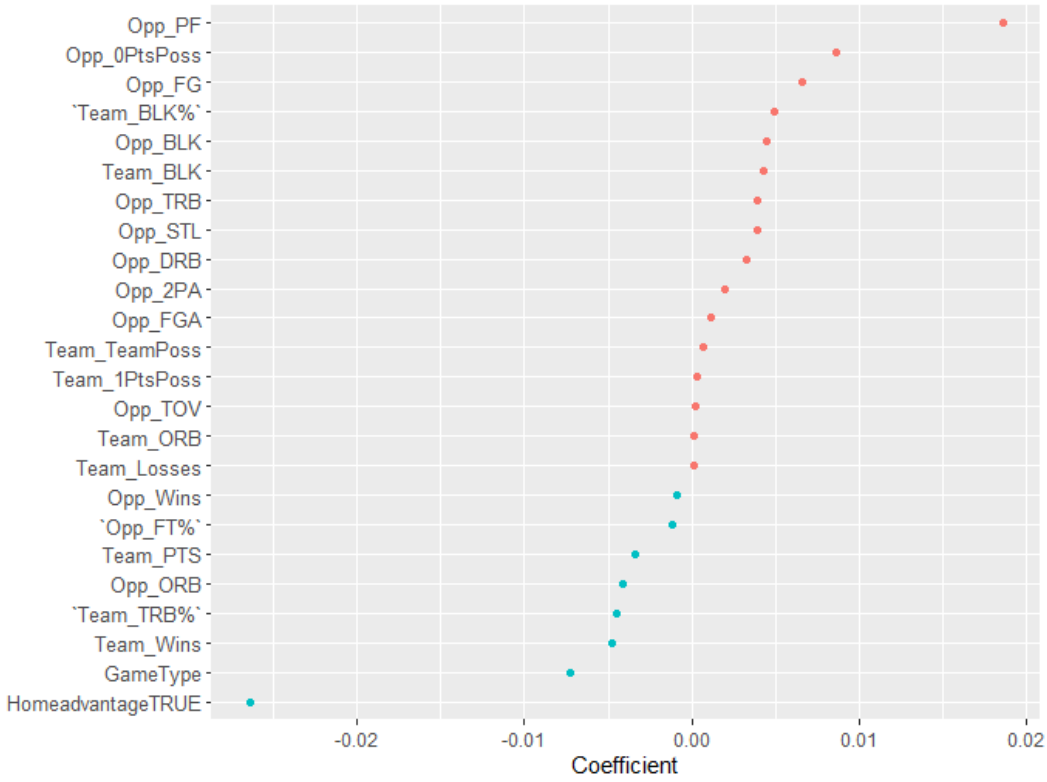


Zero-Point Plays

NBA (129 variables)

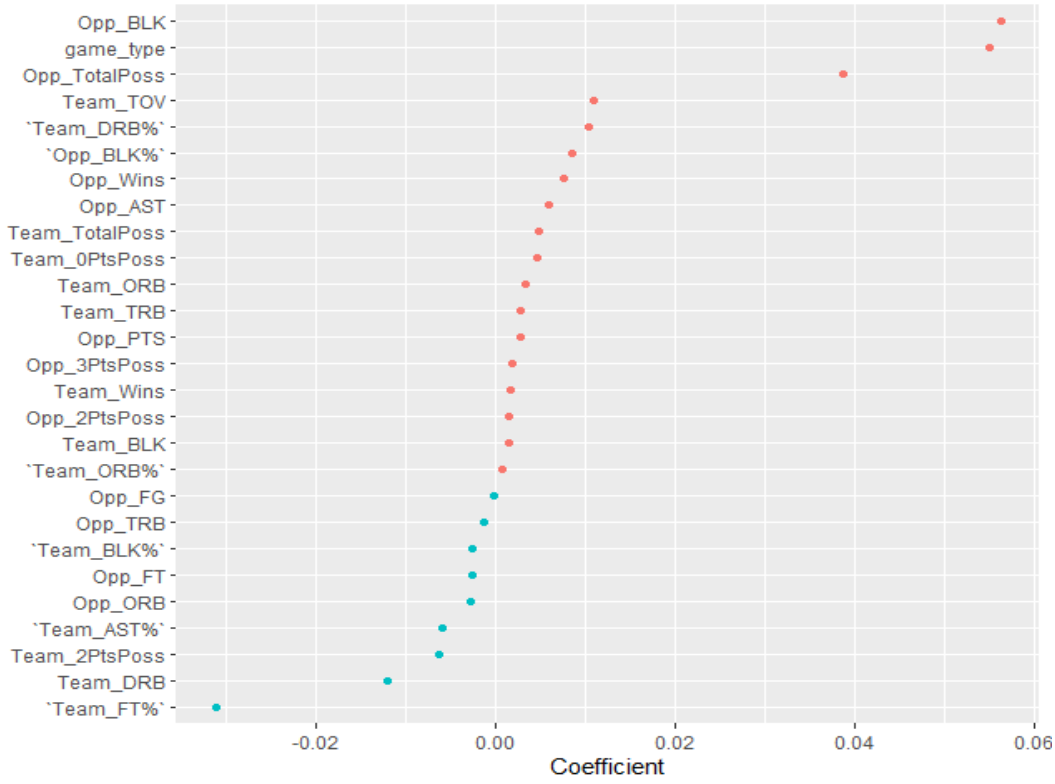


EuroLeague (85 variables)

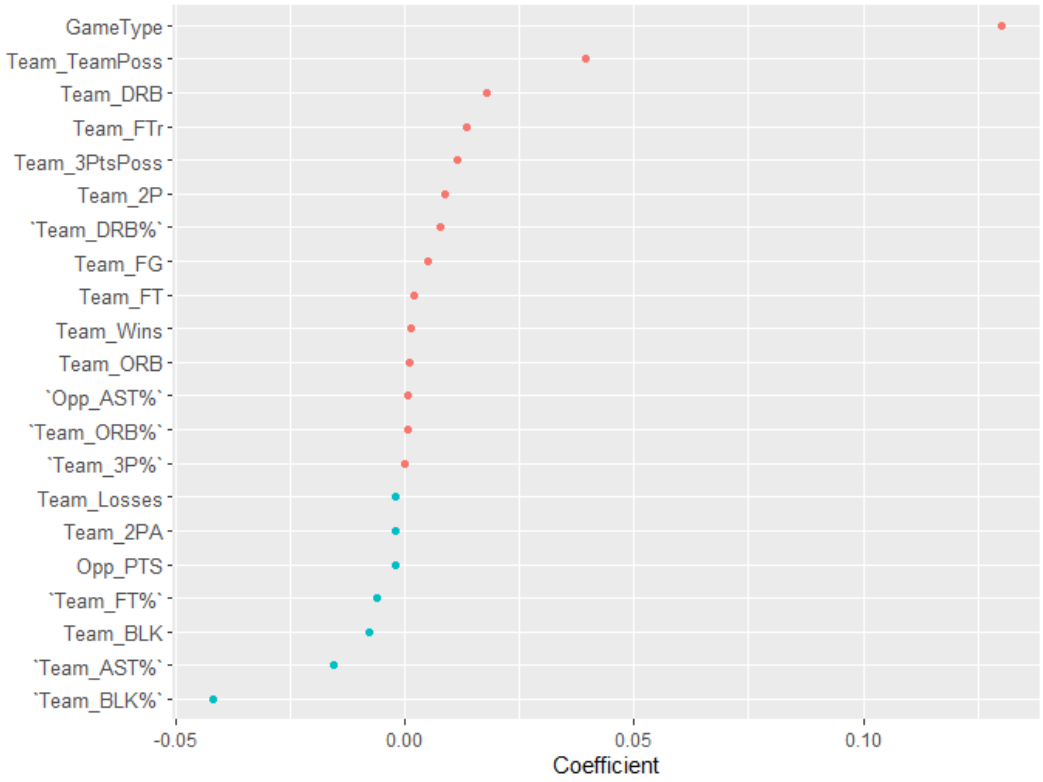


One-Point Plays

NBA (93 variables)

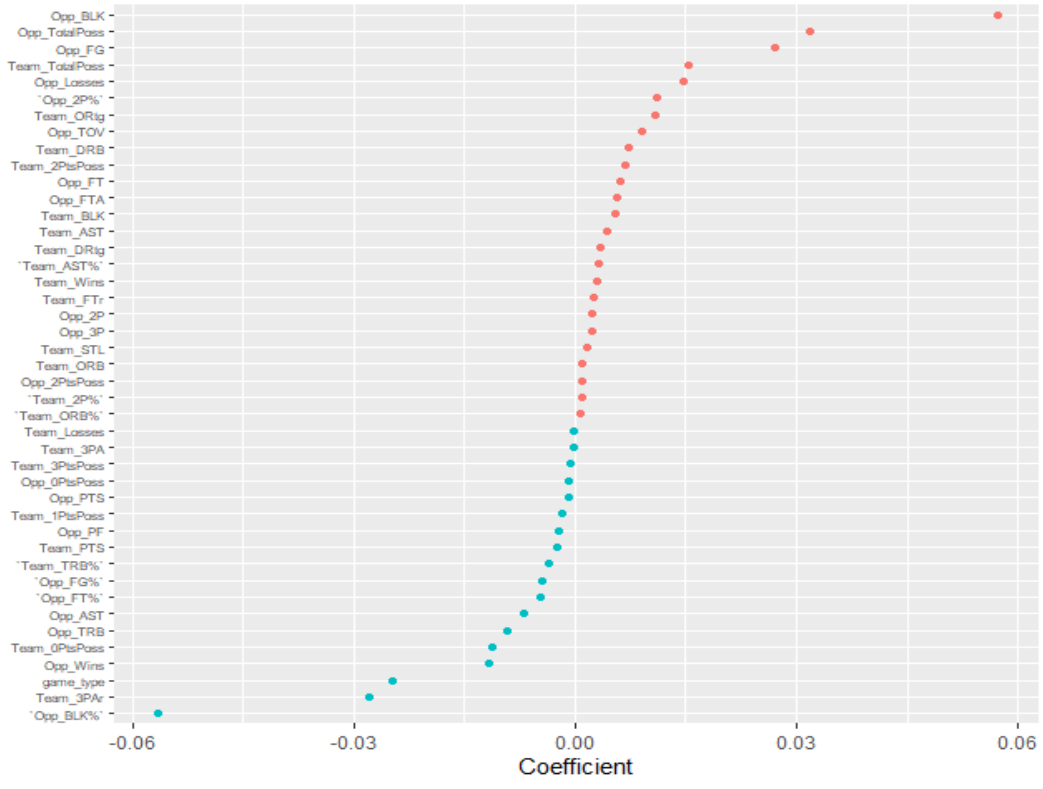


EuroLeague (83 variables)

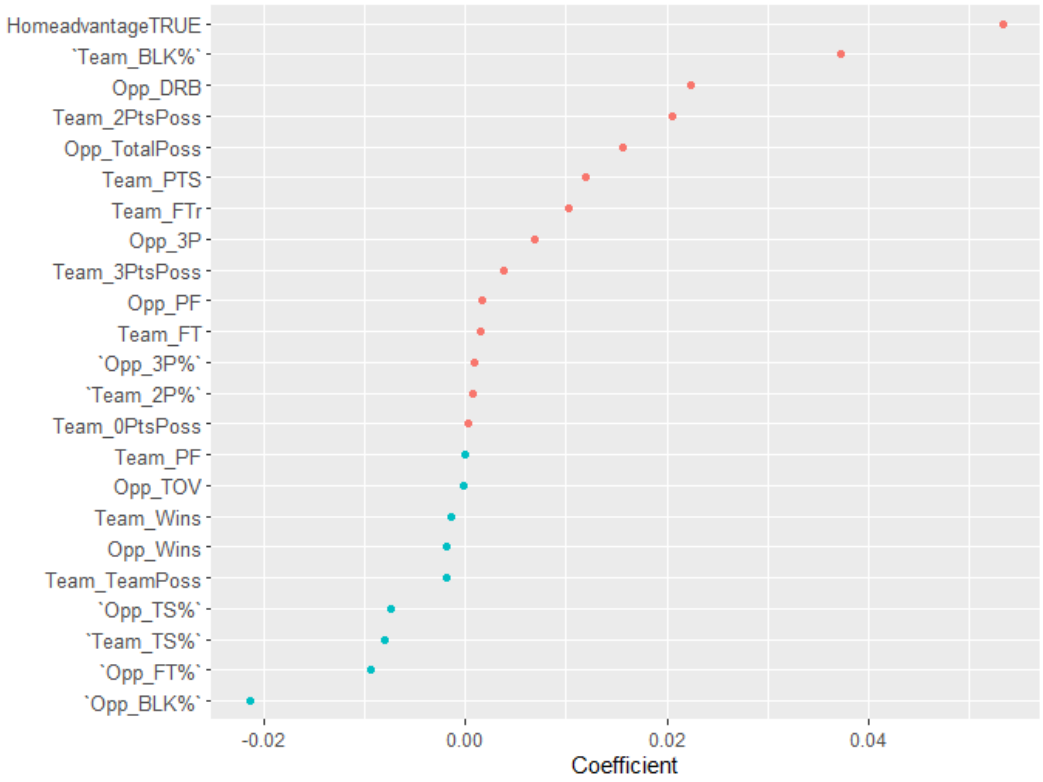


Two-Point Plays

NBA (122 variables)

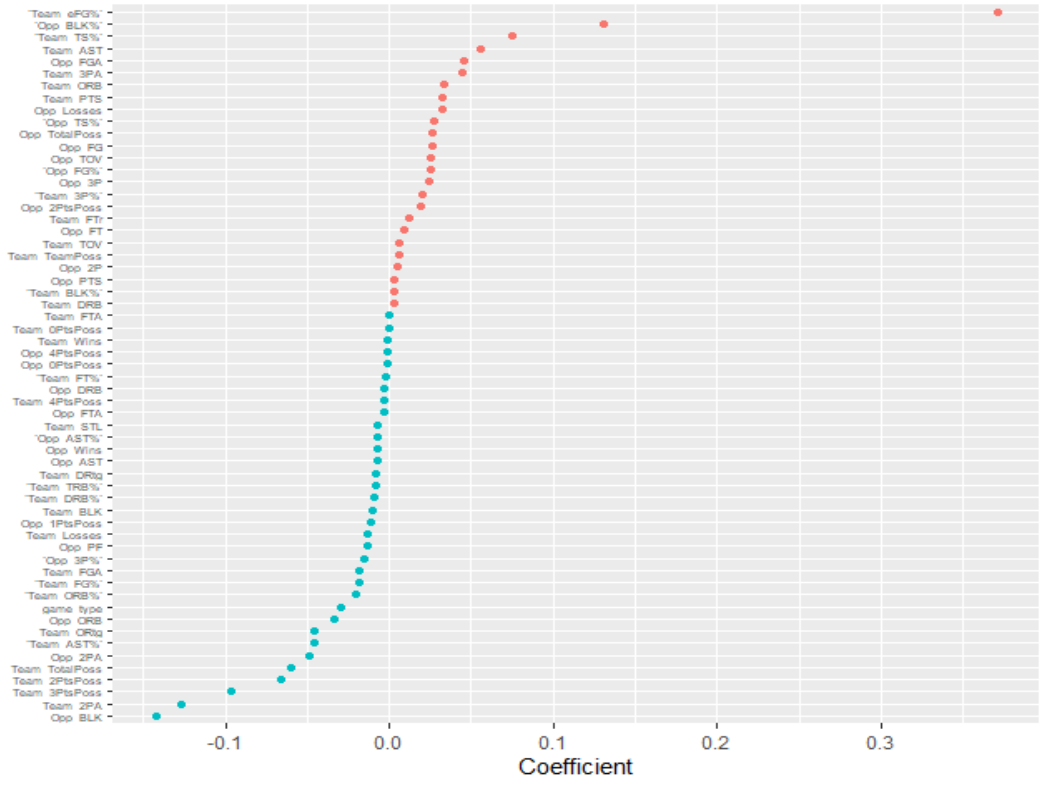


EuroLeague (117 variables)

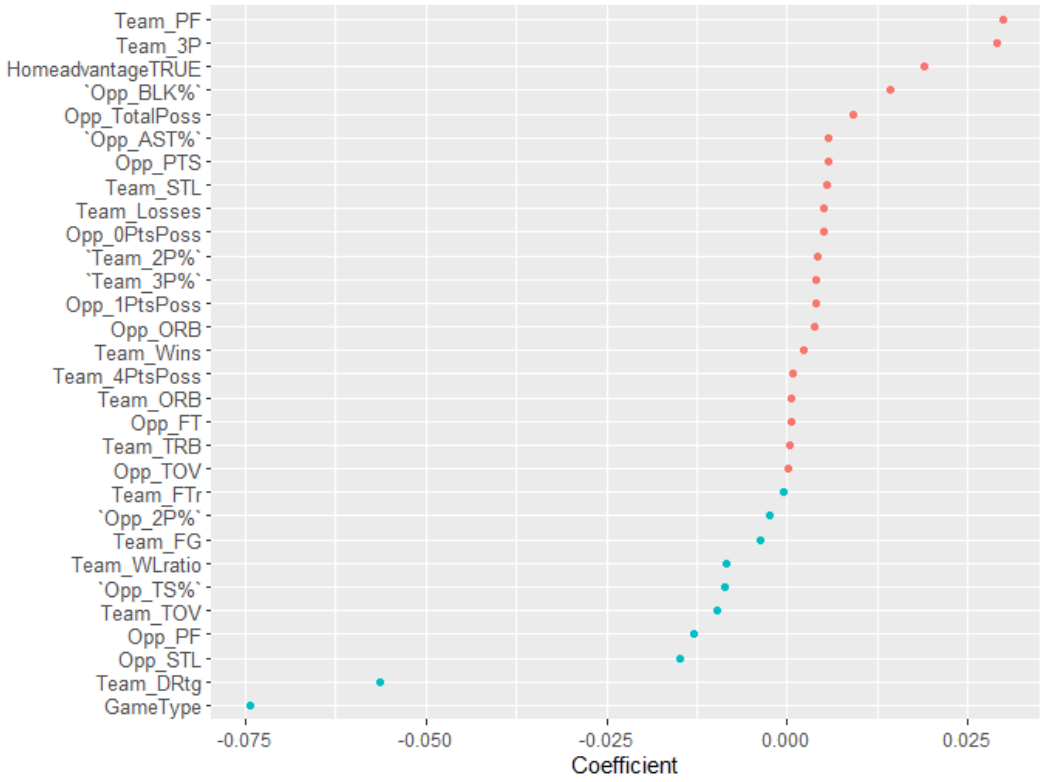


Three-Point Plays

NBA (140 variables)

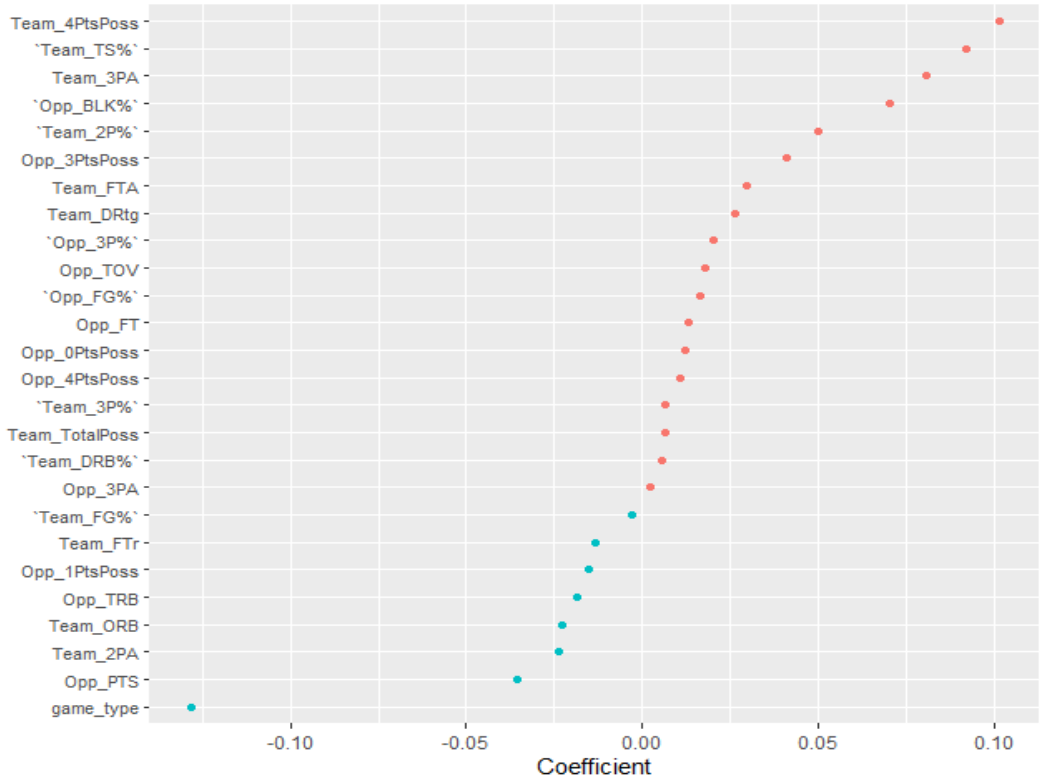


EuroLeague (87 variables)

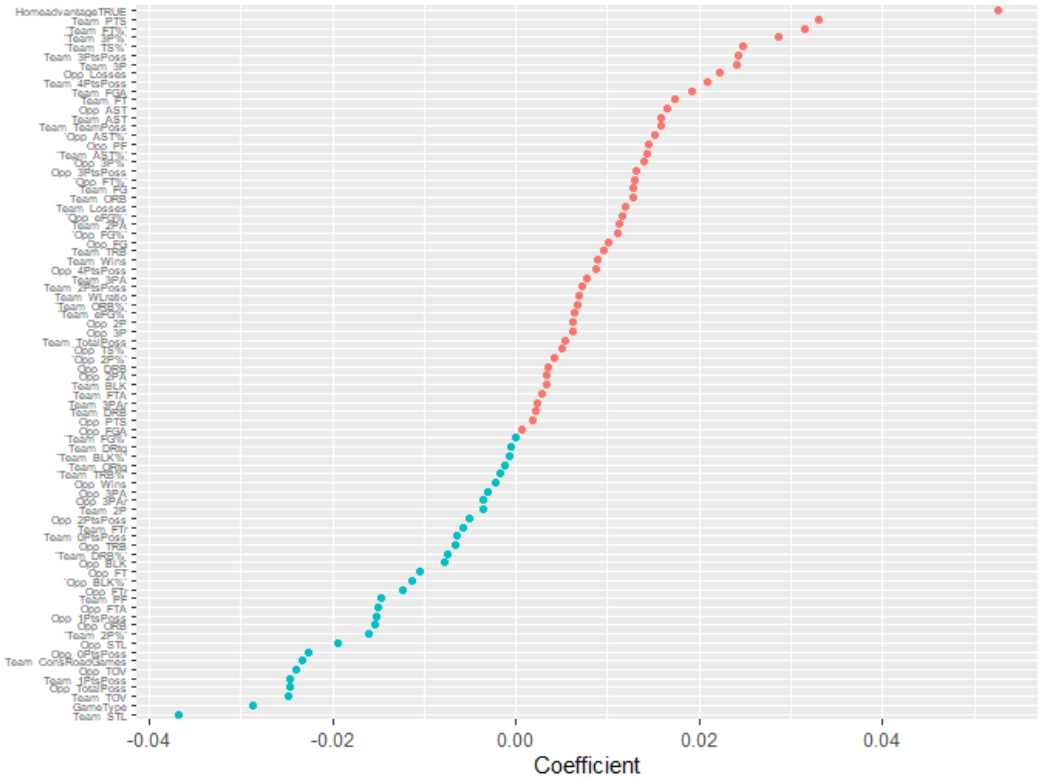


Four-Point Plays

NBA (69 variables)



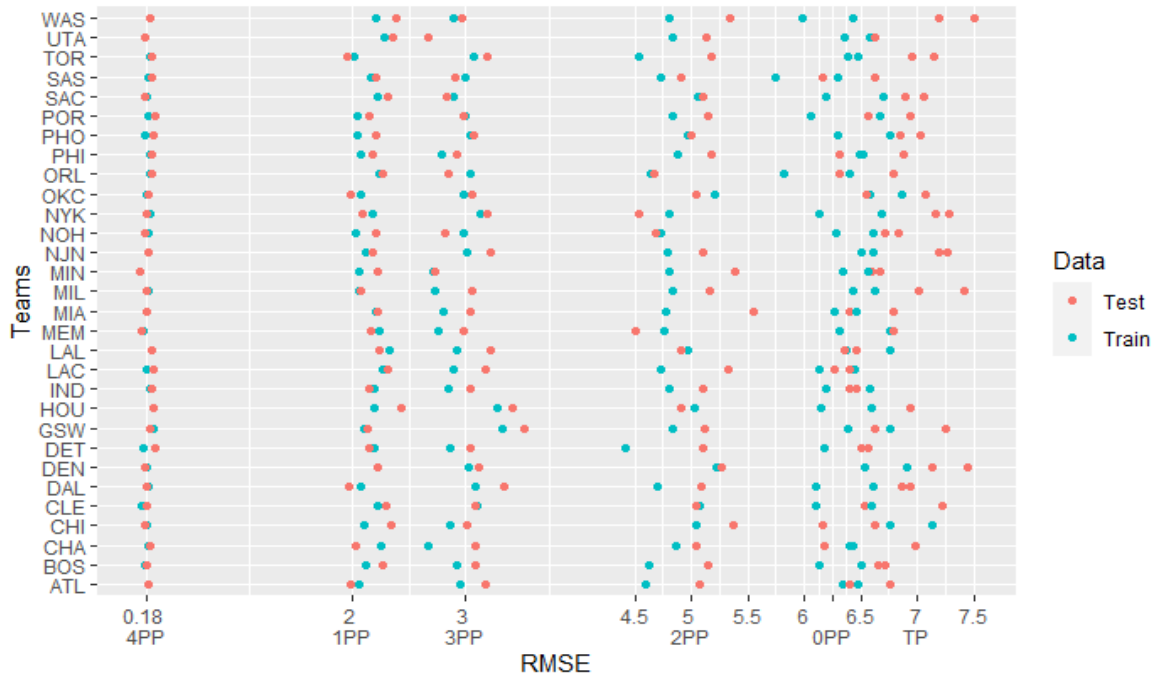
EuroLeague (all variables / 213 variables)

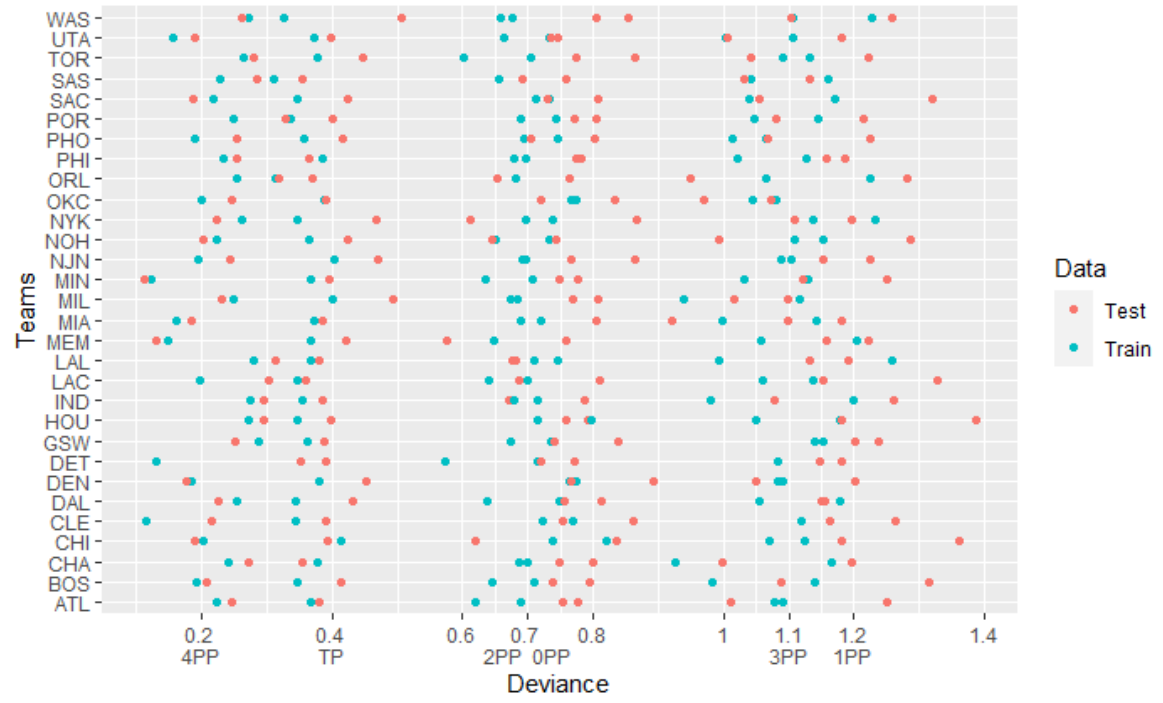


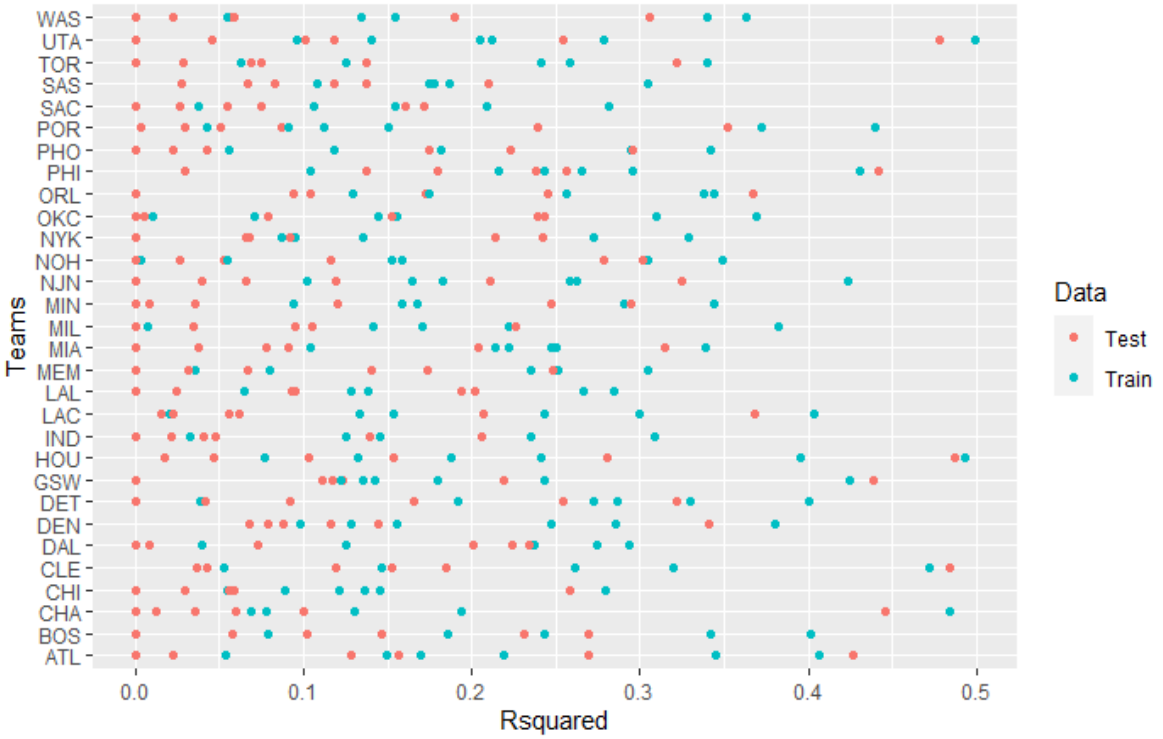
Appendix 6

Plots of the RMSE, Deviance and R^2_{DEV} of the team models with no interactions for the NBA and EuroLeague. The teams are listed on the y-axis and the values on the x-axis. The blue points are the values based on the trainings set and the red points are the values based on the test set. In figures showing the RMSE and Deviance for different NBA teams there is a clear separation between values, with values for a certain play grouped around an x-value. The type of play these values belong too is also depicted on the x-axis.

NBA

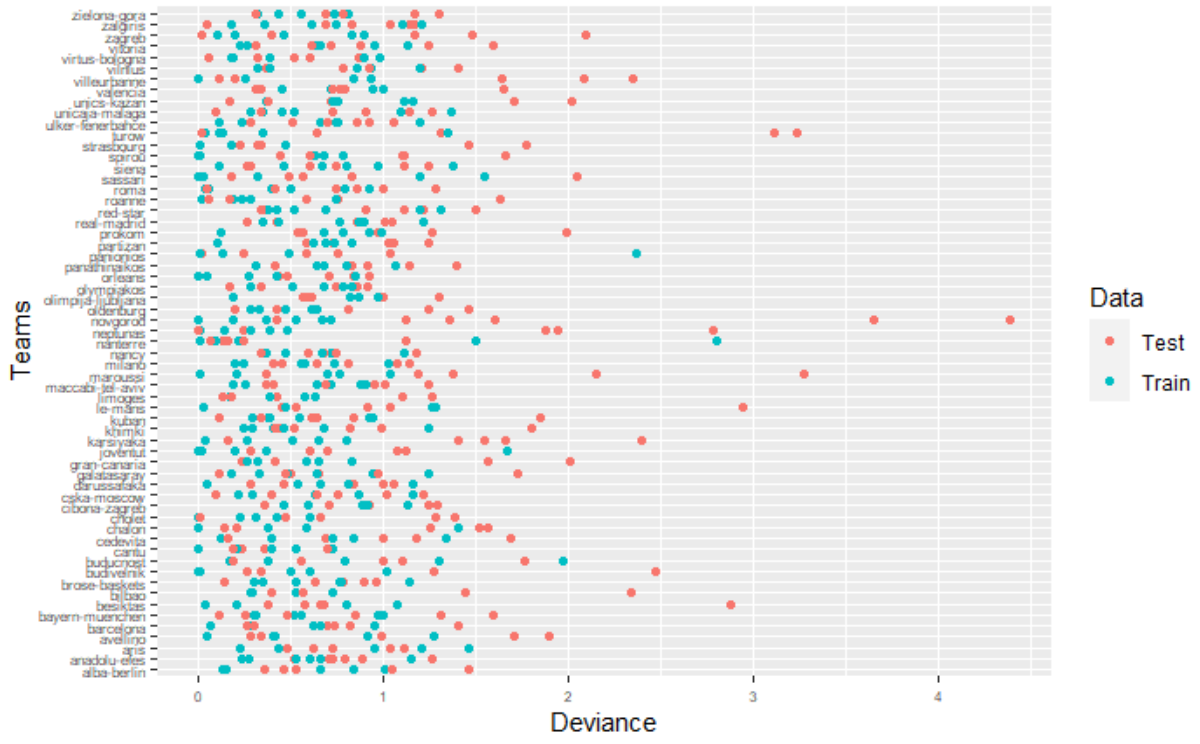


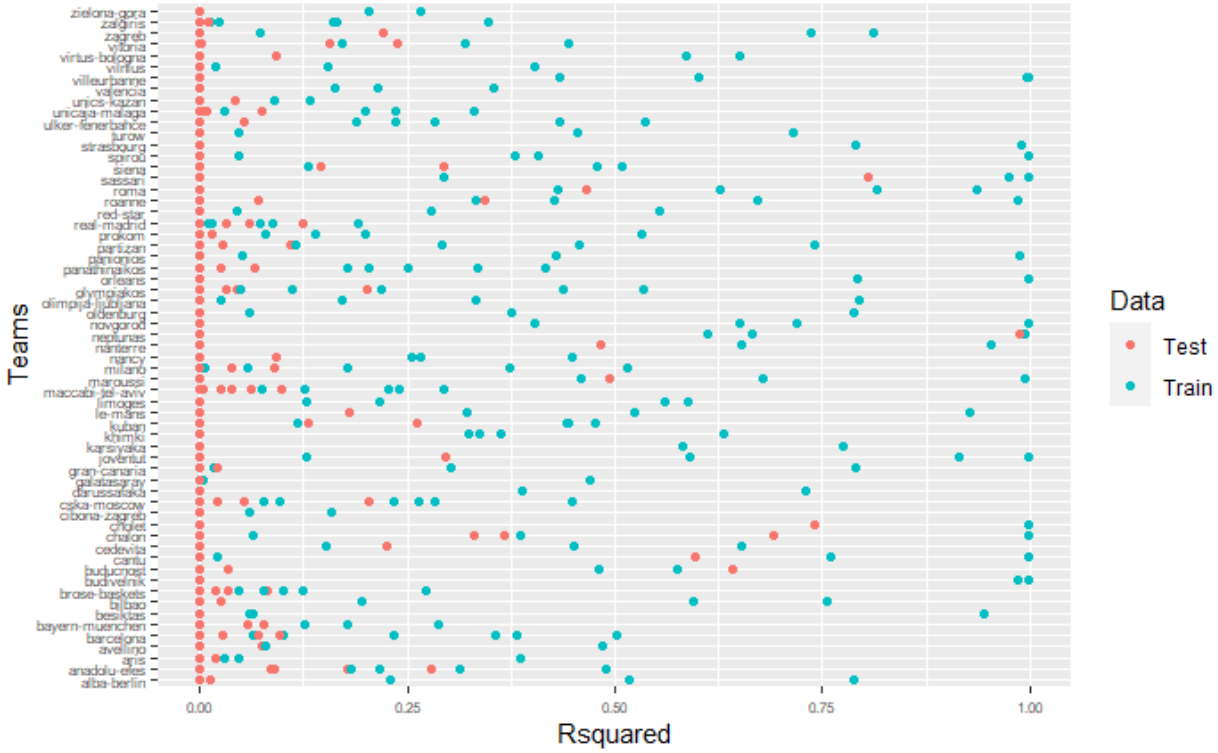

Teams
WAS
UTA
TOR
SAS
SAC
POR
PHO
PHI
ORL
OKC
NYK
NOH
NJN
MIN
MIL
MIA
MEM
LAL
LAC
IND
HOU
GSW
DET
DEN
DAL
CLE
CHI
CHA
BOS
ATL
0.2
4PP
0.4
TP
0.6
0.7
2PP
0.8
0PP
1
1.1
3PP
1.2
1PP
1.4
Deviance
Data
Test
Train


Teams
WAS
UTA
TOR
SAS
SAC
POR
PHO
PHI
ORL
OKC
NYK
NOH
NJN
MIN
MIL
MIA
MEM
LAL
LAC
IND
HOU
GSW
DET
DEN
DAL
CLE
CHI
CHA
BOS
ATL
0.0
0.1
0.2
0.3
0.4
0.5
Rsquared
Data
Test
Train

EuroLeague


Teams
RMSE
0
5
10
Data
Test
Train


Teams
Deviance
0
1
2
3
4
Data
Test
Train


Teams
zielona-gora
zalgiris
zagreb
vitoria
virtus-bologna
virtus
villeurbanne
valencia
unics-kazan
unicaja-malaga
ulker-fenerbahce
turow
strasbourg
spirou
siena
sassari
roma
roanne
red-star
real-madrid
prokom
partizan
panionios
panathinaikos
orleans
olympiakos
olimpija-ljubljana
oldenburg
novgorod
neptunas
nanterre
nancy
milano
marousi
maccabi-tel-aviv
limoges
le-mans
kuban
khimki
karsiyaka
joventut
gran-canaria
galatasaray
darussafaka
cska-moscow
cibona-zagreb
chalet
chalon
cedevita
cantu
buducnost
budivelnik
brose-baskets
bilbao
besiktas
bayern-muenchen
barcelona
avellino
aris
anadolu-efes
alba-berlin
0.00
0.25
0.50
0.75
1.00
Rsquared
Data
Test
Train

GHENT UNIVERSITY

FACULTY OF SCIENCES

Score Modelling of NBA and EuroLeague Games Using Compound Poisson Simulation

Emiel Platjouw

Master dissertation submitted to

obtain the degree of

Master of Statistical Data Analysis

Promotor: Prof. Dr. Christophe Ley

Department of Applied Mathematics,

Computer Science and Statistics

Academic year 2019-2020