The Rise of Advanced Statistics: Home Team Advantage in the NBA and the Effect of Location

Josse E. Wannet
STUDENT NUMBER: 2003334

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

dr. M. W. Klincewicz
dr. G. Saygili

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
January 2022

The Rise of Advanced Statistics: Home Team Advantage in the NBA and the Effect of Location

Josse E. Wannet

Pinpointing the precise factors causing home advantage for sports teams has eluded researchers for decades. Fans, location, biological traits, and more have all been speculated to induce the advantage. Furthermore, use of advanced statistics are becoming more prevalent in sports. This research therefore aims to determine the underlying root of the home team advantage, whilst simultaneously filling a gap in the literature by utilizing these new statistics. Through the use of decision tree models, the findings of this research can be more easily interpreted in the practical field. This research, however, finds little indication that some of the proposed factors cause home advantage. Nevertheless, the models with advanced statistics performed best, which shows promise for future research.

1. INTRODUCTION

Home team advantage has been a known phenomenon to fans, coaches, and players for quite some time. Sports leagues have even implemented some way to either favour the home team, or to ensure neutral ground for big games. For example, the Super Bowl – the NFL championship game – is always played at a neutral stadium, so no team has an unfair home advantage. Furthermore, most sports leagues with playoffs that consist of multiple games allow the higher seeded team – the team that won more games in the regular season – to have more home games. For example, in the best of 7 playoff format in the NBA, the team that won more games during the regular season plays at home in games 1, 2, and games 5 and 7 if it gets that far. This ensures that the higher seeded team has more chances to play at home, and if the teams are tied at three games each, the higher seeded team will have the additional home advantage in the win or go home deciding game 7. This gives more meaning to the regular season for the sports leagues that utilize these postseason playoffs/tournaments. However, quantifying what exactly causes the home advantages has proven to be difficult. Research has investigated the effect of crowd noise, travel time for the opponent, location of the home team, and more.

Recent literature that focusses on analysing the home team advantage often use regression models to support their research (e.g., Anders and Rotthoff (2014); Areni (2014); Chang, Ran, and Smith (2021); Van Damme and Baert (2019); Willoughby and Becker (2014)). It is however notable to see the lack of use of more 'advanced' statistical methods, i.e., machine learning algorithms. It seems that Harris and Roebber (2019) are so far the only authors to have used such a model; being an artificial neural network (ANN). Their research used team season statistics and the ANN to investigate what contributes to the home advantage. They decided to use an ANN, as, according to them, "neural networks are preferred when non-linearities in the data may be important and we do not wish to specify their structure" (p. 3). They verified their claim by showing

that the performance of the ANN is better than that of multiple linear regression models (R2 = 0.7 vs R2 < 0.5), regardless of the inputs. To thus expand the current literature on home advantage in sports and its causing factors, this current research will aim to investigate the causes for home advantage by ways of using a more sophisticated machine learning method, namely decision tree models. In order to achieve this aim, the following main research question is established:

> *To what extend does home team advantage occur within the NBA?*

To furthermore substantiate the research question, a few subquestions have also been formulated. This research intends to build on the work of Harris and Roebber (2019), where they mention in the discussion and limitations of their paper that it would be worthwhile to investigate the effect of blocks, steals, and fouls into future analysis. They excluded these statistics from their model. Their reason for exclusion is not mentioned. The importance of these statistics for the home team advantage is, however, often underlined in other research (Goldman and Rao (2012); Leota et al. (2021); Nevill, Balmer, and Williams (2002); Roeder (2017)). As such, to further support the main research question and to build upon the work of Harris and Roebber (2019), the following sub question is established:

> RQ1 *To what extend do blocks, steals, and fouls contribute to the advantage of the home team?*

Furthermore, there has been a rise in the use of so-called 'advanced statistics' within the NBA (Gobikas, Radu, and Miklovas 2020). According to the Houston Press, since 2007, Daryl Morey – then General Manager of the Houston Rockets – started using these advanced statistics as a key aspect of player evaluation (Friedman 2007). These stats include for example the offensive rating of a team per 100 possessions (so normalized for pace to allow for better comparisons), pace (possessions per 48 minutes – length of game), true shooting percentage (a "better" indicator of player shooting efficiency), and more. These statistics might be a better representation of on court play, and as such, should be included in the analysis. More, recent literature has not yet adopted the use of the statistics, so this research aims to fill this gap.

> RQ2 *Seeing the rise in popularity of 'advanced statistics' within the NBA, are these newly adopted statistics good predictors of home team advantage?*

Lastly, as mentioned previously, explanations of what causes or enables the home team advantage vary. In addition to using a more advanced machine learning algorithm and the inclusion of blocks, steals, and fouls statistics, this current research will also use location of an NBA team as possible explanatory variables for the home team advantage. Location as explanation for the home team advantage is prominent in literature on this topic. In the literature, location is often subdivided into two categories: travel and altitude. McHill and Chinoy (2020) examined the impact travel for the away team has on overall team performance, and how it could help explain the home team advantage. They used the natural experiment of the 'bubble restart' of the 2019/2020 NBA season. Due to COVID-19, the final eight games of the regular season and the whole playoffs were played in an isolation zone at Walt Disney World in Florida. Thusly, the research had data on games played when teams had to travel (pre COVID-19) and when 'away'

teams had no travel time (the bubble restart). The authors found that winning percentage only significantly differed for teams when they had to travel across time zones, mostly when traveling westward.

Secondly, the effect of elevation/altitude of the city a NBA team resides in, has shown to have an effect on the home advantage of teams. Paine (2013) found that the average margin of victory for the home team is 3.2 points (data from season 2004/2005 till around March 2013). Meanwhile, the Denver Nuggets and the Utah Jazz have a significantly higher average margin of victory at home of 5.5 and 6.2 points respectively. This could be linked to the two teams being located in the two highest altitude cities in the NBA. The thinner air at these high-altitude locations could prove problematic for the oxygen intake for the away team players, causing quicker fatigue. The home team players are more acclimated to the thinner air, as they spend more time at these altitudes. The same result was found by Lopez, Matthews, and Baumer (2018), who saw that all Denver located teams had the largest home advantage (MLB, NBA, NFL, 7-th highest in the NHL). They too speculated that the acclimation to the high altitude was the reason for this consistent advantage across sports.

Finally, this research on location impact on the home advantage also includes rivalry into the analysis. The NBA is divided into two conferences (East and West), which are further divided into three divisions each. The division of conference and division is based on distance between the teams. Importantly, teams in the same division and conference will play each other more often throughout the season, which can cause rivalry to develop between these teams in close proximity. Neave and Wolfson (2003) have shown that rivalry can have an effect on home advantage, by discovering that testosterone levels of soccer players playing at home were higher when playing a perceived 'extreme' rival than a 'moderate' rival. They state that the increase in testosterone could possibly lead to a better performance, and thus an increase in home advantage. As such, rivalry by proxy of location could be an explanatory variable for the home team advantage.

This results in the third and final sub research question:

RQ3 *To what extend does location affect the home team advantage? Location is defined as the absolute distance between teams, time zone difference between teams, and division/conference rivalry.*

2. RELATED WORK

Extant research on home team advantage focusses on several factors. For this research, the sections below will explore the two main factors associated with home team advantage that are included in the current research. The following sections furthermore show that recent research mainly use regression models to obtain results, which is further proof that the use of a more advanced analytical model could provide a new foundation for future research.

2.1 THE HOME CROWD

Intuitively, the home crowd is often thought of as the biggest factor influencing home advantage in sports. The sounds of a passionate home crowd can support and encourage the home team, whilst simultaneously discouraging the away team. Over 40 years ago, Schwartz and Barsky (1977) already established that part of the home advantage

is caused by the social support of partisan fans. Their research showed that fans help boost the offensive performance of the home team, rather than ensuring visiting teams perform poorer on the defensive end. In their study, they also found that the simple fact of playing on 'your own court', was as much of a determinant of the game's outcome as was the quality of the teams in the ballgame. Their research is one of the first occurrences of analysis on home advantage in sports, and since then, the scope of home advantage has increased in the literature.

The effect of home advantage is especially prominent in the NBA, as the indoor arenas result in the fans being closer to the players, which provide the players with extra motivation through the strengthened noise (Schwartz and Barsky 1977). Building on the work of Schwartz and Barsky, Greer (1983) performed a quasi-experimental field study where a behaviour observation technique was used to observe home and visiting male basketball teams at two large state universities, to investigate the effect of sustained spectator protest on performance of both the home and visiting team. Through the use of a hierarchical multiple regression model, these "postbooing" periods were compared to performance during periods of normal spectator conditions. Analysis showed that spectator protests would lead to slightly better performance of the home team and significant declines of the visiting team. Greer concludes that crowd noise can inhibit the performance of the visiting team.

Familiarity of the stadium has also proven to be a cause of home advantage. Pollard (2002) found a significant decrease of the home advantage for teams moving to a new stadium within the new city. In this study, home advantage was calculated as the number of home wins compared to the total number of wins in a given season, and expressed as a percentage. Their study of 37 professional North American teams found an average reduction in home advantage of 23.50% (17.50% for the 17 NBA teams included in the study). This decrease in home advantage was found to be significant (t = 2.39, P = 0.011), with significance tested using a standard paired t-test. The familiarity with the home stadium due to being able to train and play there more often, can help a player become more accustomed in their rhythm, preparation, and possible even help shooting perception. Anecdotally, Steph Curry – one of, if not the, greatest three-point shooters in the history of the NBA – has notably shot the three-point ball worse when visiting the Los Angeles Lakers than when visiting the Los Angeles Clippers, even though the two L.A. teams both play in the same stadium: Staples Centre. In 35 total games, Curry has a total field goal percentage of 41.8% and a three-point field goal percentage of 34.9% against the Lakers, whereas these same averages are respectively 50.1% and 46.7% against the Clippers. A possible explanation for this marked difference is that the Lakers use different lighting during their games, giving it a darker, more dramatic, and theatre feel. The Clippers use 'regular' lighting during games, lighting that Steph Curry is used to from his home stadium. As such, familiarity with a stadium could possibly influence a player's performance, which in turn impacts the home advantage. Literature thus shows that playing at home in front of the home crowd in attendance is vital in any analysis on the home team advantage. Ergo, this current research will implement statistics on attendance to help answer the main research question: *'To what extend does home team advantage occur within the NBA?'*.

Further research by Goldman and Rao (2012) showed that players playing at home get better at 'hustle plays' – effort-based plays – such as offensive rebounding (when the offensive team rebounds their own missed shot attempt, thus allowing for another offensive possession and possibility to score) during pressure packed moments. The authors analysed over 1.3 million possessions using regression analysis, to come to the conclusion that the home team secures more offensive rebounds, which gives them an

increased edge over the visiting team. They conclude that this shows that the crowd can inspire a home team to put in more effort, which can lead to better performance. The result of this research was underlined by Leota et al. (2021). The authors performed a study on the effect of the home crowd by using the 2020/2021 NBA season, where approximately 50% of the games were played without a crowd due to the COVID-19 pandemic, with the other games being played with a highly reduced fan attendance. Using mixed linear models, they show that the home team out rebounds (effort-based plays) the visiting team by as much as 2.28 rebounds on average when a crowd is present. This is in significant contrast to games played without crowds, where the home team fails to out rebound (average rebound differential = -0.41 rebounds) the visiting team.

Similarly, whilst blocks and steals of a team are related to the defensive schemes of an NBA team, they are also indicators of the effort a player puts into the defensive end during an NBA game. Whilst the defensive scheme can put a player in the right position, the player still needs to put in the effort to try to steal or block the ball. Blocks and steals can thus be considered as hustle plays, where the aforementioned literature has shown that these increase at home. This validates the assumption Harris and Roebber (2019) made in their discussion and underlines the importance of examining these statistics and their relationship with the home team advantage.

Furthermore, this research also includes fouls into the analysis on the home advantage. Literature has shown the effect the home crowd can have on referees and the fouls they call during a game. With the use of logistic regression analysis, Nevill, Balmer, and Williams (2002) found that there is a significant effect of fewer fouls being called against the home team, which they attributed to crowd noise. The authors looked whether decisions by qualified referees could be influenced by the noise of the home crowd. Participating referees were asked to observe 47 videos of incidents. Twenty-two referees observed these videos with audible crowd noise, whilst 18 referees viewed the videos with no sound. They were then asked to indicate if what they saw on video was either a foul or not a foul. More, they were also asked to indicate if it was a foul for the home or away team. If unsure, the referees were allowed to fill in 'uncertain' as a response (in a live game, 'uncertain' would result in 'no foul'). The noise condition referees were found to be more uncertain in their decision, and awarded 15.5% fewer fouls against the home team than the referees who viewed the videos in silence. The significance of this result was tested using backward elimination (according to Draper and Smith (1981)) in which importance of a variable is assessed by the 'change in deviation' (X^2) that results from dropping the variable. Removing the noise factor group from the final model resulted in a significantly large change in deviance ($X^2 = 3.875$, $p < 0.05$).

Another study by Roeder (2017) analysed referee decisions during the last two minutes of close games (defined as games being within 5 points or fewer) using the NBA's database of calls. These specific circumstances were chosen, as since March 2015, the NBA itself began assessing referee decisions of games in the aforementioned circumstances. This ensures more validity of assessment of correctness of calls. Using this data, Roeder (2017) categorized the calls into three categories: correct calls, incorrect no-calls, and incorrect calls. They found that away teams have more incorrect calls called against them and the home team benefits more from incorrect non-calls (i.e., calls that should have gone against the home team but were not made). These studies clearly show the effect the home crowd can have on referees, which in turn affects the fouls called during a game. Fewer fouls called against the home team, mean fewer free throws for the visiting team and a smaller likelihood of a crucial home team player to foul out

of the game. These two factors can lead to a higher home team advantage, and thus foul statistics are included in this research.

Together with the addition of block and steal statistics, the inclusion of fouls aims to examine the home team advantage through answering the first sub research question: *'To what extend do blocks, steals, and fouls contribute to the advantage of the home team?'*.

2.2 LOCATION

As mentioned in the introduction, location of the home team has often been speculated to be a (partial) cause of the home advantage. Location of the team is commonly split into various impacting factors, of which elevation and travel for the visiting team are of interest for this current research. Furthermore, travel for the visiting team encapsulates multiple factors. Of importance for this research is distance between NBA teams, and travel over time zones for the visiting team.

A regression analysis by Goumas (2014) found that the home advantage increased by a relative 20% for each time zone that was crossed by a visiting team (significant results with $p < 0.001$). Similarly, Oberhofer, Philippovich, and Winner (2010) used data from the German Football Premier League and a Poisson regression model to show that performance of a team decreased as distance to the venue where the game is played increased. The same was found in one of the earliest studies on distance and home advantage, when Snyder and Purdy (1985) found that when a visiting team was within 200 miles (322 kilometres) of the home team, the home team winning percentage was 58.8%. When the visiting team had to travel more than 200 miles, the home team winning percentage skyrocketed to 84.6%.

Additionally, the distance between two competing teams also has an effect on the attendance during games. Winfree* et al. (2004) investigated the effect of the distance to the nearest competing professional sports team within the MLB (baseball). With the use of non-linear generalized least squares estimation, they concluded that two teams that are closer together, will have lower attendance than two teams farther apart. A one mile (1.6 km) increase in distance to another MLB team increases the attendance by around 1544 fans. Furthermore, the addition of a team in the vicinity would reduce attendance by an additional 126,500 fans. The authors mention an interesting example of the Dodgers sharing Dodger Stadium with the expansion Angels, which cost the Dodgers around 1.76 million in 1963 dollars (that's close to 16 million dollars today).

These studies show the importance location can have on the home advantage, both in terms of pure travel effects for the visiting team, as well as effects on the home crowd; more support for its inclusion into this research.

The effect of elevation of the home team location has been researched to a much lesser extent in regard to its effect on home advantage, yet there are some studies that show its importance. Lopez, Matthews, and Baumer (2018) and Paine (2013) were already mentioned to have found a connection between the relatively high-altitude cities of Denver and Utah, and the increased home advantage of the teams playing in those cities. Further, McSharry (2007) and Pollard and Armatas (2017) have shown a significant association between home advantage and altitude of the home team location in studies analysing football matches. McSharry (2007) showed that an increase of altitude difference of 1000 metres would, on average, result in a goal difference of half of a goal for the home team, whereas Pollard and Armatas (2017) showed a 0.115 point advantage for the home team for each 1000 metres above the visiting team's altitude.

This current research will therefore further substantiate the literature on the effect of altitude and home team advantage, through incorporating altitude data in the analysis.

Lastly, one of the important aspects fans always think of when it comes to the home team advantage, is the rivalry the home team might have with a visiting team. Rivalries exist within all sports, be it same city football derbies such as the "Old Firm" (Rangers vs. Celtic, Scotland), "Merseyside" (Liverpool FC vs. Everton FC, England), or rivalries such as the Chicago Bears vs. the Green Bay Packers (NFL, United States); a century long rivalry due to being matched up in the same division for all those years, rivalries have existed for quite some time. Most of these rivalries exist and were established due to close proximity between two teams within the same sport. This sort of rivalry can also be found in the NBA, where as recently as 2018 one of the top NBA players requested a trade from the San Antonio Spurs, a Western Conference team. His preferred trade destinations was the Los Angeles Lakers or the Los Angeles Clippers, both also Western Conference teams. A rumour then circulated that the Spurs would rather trade the player to the Eastern Conference, even if that means that they would "take 75 cents on the dollar", Feldman (2018) reports on the NBCSports website. This was presumably caused by the Spurs' rivalry with Western Conference teams, and the L.A. Lakers specifically due to some heated playoffs matchups over the years. This goes to show that rivalry exists even when teams are not necessarily proximate in absolute distance (San Antonio – Los Angeles is 2186 km).

Research on rivalry is, however, severely lacking. The study of Neave and Wolfson (2003) and the effect of rivalry, testosterone levels, and home advantage, as mentioned in the introduction, is one of the few. As argued above, extreme rivals often are teams that are close to each other in distance, but could simultaneously be teams that face each other often. Within the NBA, that would be teams that are in the same division or conference.

This current research therefore aims to fill this gap in the literature by also including division and conference rivalry in the analysis of location effects on the home advantage. In this manner both possibilities that can create a rivalry – distance between teams and playing each other often – are included in this study.

To conclude, the factors of distance, travel, and rivalry are examined in this research, with the aim to answer the third sub research question: *'To what extend does location affect the home team advantage? Location is defined as the absolute distance between teams, time zone difference between teams, and division/conference rivalry.'*.

3. METHODS

For data handling and analysis, several Python packages were used. Most importantly Pandas (McKinney et al. 2010), NumPy (Harris et al. 2020), and GeoPy[1] were used for data handling and processing. For modelling and analysis purposes, Scikit-learn (Pedregosa et al. 2011), and statsmodels (Seabold and Perktold 2010) were used, and finally Seaborn (Waskom et al. 2017), dtreeviz[2], and Matplotlib (Hunter 2007) were instrumental in visualization purposes.

1 https://geopy.readthedocs.io/en/stable/
2 https://github.com/parrt/dtreeviz

3.1 DATA

The data used in this research is obtained from Basketball Reference[3], a freely available website using data provided by SportRadar, which is the official statistics provider of the NBA. This data source is highly reliable and used in most research regarding the NBA. The data includes the game logs of each of the 30 teams, ranging from season 2009-2010 to season 2020-2021. Season 2011-2012 was excluded from the analysis, due to it being a lockout shortened season. The lockout was caused by disagreement between team owners and players on the new Collective Bargaining Agreement. This 'abnormal' season could have a unknown effect on the data, and to stay within scope of the research, data from this season was not used. Season 2019-2020 is only partially included, as the last few games were played in the 'Orlando bubble', where no fans were in attendance at all. There has been plenty of research on the effect the bubble and COVID-19 had on the teams (e.g., Ehrlich and Ghimire (2020); Fischer and Haucap (2020); Higgs (2021); Loures, Shikida, and Fernandez (2021); Price and Yan (2021)), so these games are also left out of the analysis. Additionally, during season 2020-2021 most teams had no fans or reduced crowds in attendance due to COVID-19 regulations. These differences in attendance can provide useful analysis for the effect of the home crowd on home advantage and are thus not excluded from the data. Lastly, 20 games were excluded from the data set, as these games were part of the NBA's expansion into the international scene, such that these games were played in London, Mexico City, or Paris. The 'home team' therefore does not actually play at home, and should thus be excluded from analysis.

Data from the game logs includes both 'basic' statistics, as well as the 'advanced' statistics. An explanation of these statistics – provided by Basketball Reference itself – and of all variables in the data set can be found in Appendix: Table 7. Furthermore, the game logs included these statistics for both the home team as well as for the visiting team. As such, it is possible to see, for example, the shooting percentages or rebound percentage of the visiting team. If those are low, it could mean that the home team is defending well. This could be explained by the effect the home crowd can have on these hustle plays, which in turn helps answer the research questions.

Furthermore, attendance data was also obtained through Basketball Reference. Arena capacity data was obtained through Wikipedia[4], which obtained capacity data from current arenas from the NBA itself, whilst capacity of former arenas was obtained through various other references.

Coordinates and elevation data of the arenas were obtained by use of data from Lewis Pipkin who posted it for free use on GitHub[5]. They obtained this data through Google Maps. Obtaining this data personally was not possible, due to Google implementing their API system behind a paywall.

Finally, time zones of the NBA teams were collected based on the cities the teams are located in.

3 https://www.basketball-reference.com/
4 https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_arenas
5 https://github.com/lewispipkin/NBA

3.2 DATA PREPROCESSING

Basketball Reference allows for data extraction to CSV files on their website. Therefore, all needed data was separately downloaded, after which the necessary files were appended and merged as needed. From this, full game log statistics of both basic and advanced statistics for all included seasons were merged into a final data set. From this data set, games played between July 30th and August 14th of 2020 were filtered out, as these games were played within the NBA Bubble and should be excluded (see section 3.1).

Then, using the database of Basketball Reference again, attendance data was also merged into a single file. After renaming the columns and the team names from full names to shorthand, attendance data was merged with the game log data set. As such, the final data set showed the attendance for each individual game, allowing analysis of the effect of crowd size and attendance on game outcome and the home advantage.

To calculate the distance between NBA teams, the geodesic function of Python package GeoPy was used. The geodesic distance is the shortest distance between two points on a curved surface, such as the Earth. This is analogous to the distance of a straight line on a plane surface. To calculate this distance, GeoPy uses the algorithm established by Karney (2013). From Pipkin's data set, the latitude and longitude data of the NBA arenas were zipped into a list. The geodesic function was then applied to these data and arranged in a square data frame. The result of this can be found in Appendix: Figure 4. This figure shows the corresponding distance difference in kilometres between arenas of each NBA team. Furthermore, to answer part of the third sub research question, division and conference rivalry were added to the data by inserting dummy variables. If both teams are in the same conference, a 1 is inserted in the conference column. If both teams are then also in the same division, a 1 is inserted in the division column. If the teams are not in the same conference or division, a 0 is added to either column.

A simple calculation of elevation differences gave the resulting data frame that can be seen in Appendix: Figure 5. Before this data was further used, it was converted from feet to metres. In the final data set, the elevation difference between home team and visiting team was included, as well as the absolute elevation of the home team. It is therefore possible to see the influence of altitude in and of itself, as well as the effect of change in altitude between playing at home or away. Analysis is expected to show if elevation itself is a predictor of home advantage, or if a change in elevation is needed to be a factor in increasing home advantage.

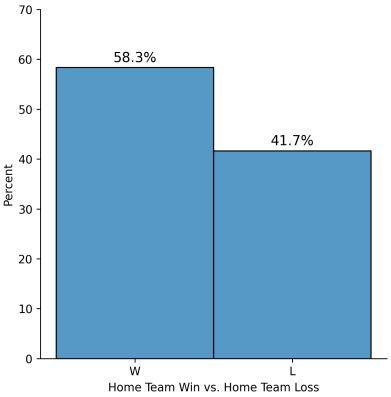
Then, time zone differences were added to the teams. Initially, these data were a concatenation of strings of the two time zones of the teams. For example, a home game by the Atlanta Hawks (Eastern Time Zone) and the visiting Golden State Warriors (Pacific Time Zone) would be: "ETPT". These strings were then converted into a numerical value, based on the difference in time zones. In this case, there is a three time zone difference between the Eastern Time Zone and Pacific Time Zone, so the corresponding value is 3. These differences can be seen in Appendix: Figure 6.

Lastly, the season in which a game was played was also added to the data, for descriptive statistic and visualization purposes.

The distance differences, elevation differences, absolute elevation of the home team, and time zone differences, were then merged with the data set that already contained the game log and attendance data. This final data set, consisting of 13100 game instances and 65 variables, is used for analysis.

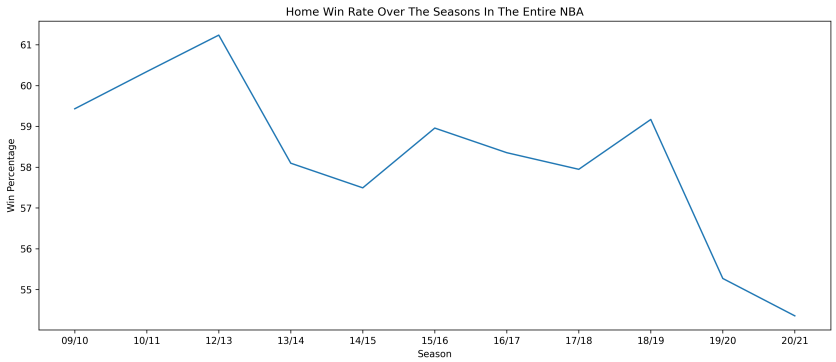
3.3 DESCRIPTIVE STATISTICS

Figure 1
Overall Home Win vs. Loss



A table of descriptive statistics of all variables in the data set can be found in Appendix: Table 8. An initial look at this table already shows a higher average score for the home team over the visiting team: mean of 105.46 and 102.89 respectively. With the data set consisting of 13100 games, this difference (2.57) is quite large. Figures 1 and 2 further show the overall home win rate in the NBA and the home win rate over the seasons included in this research. With 58.3% of the games in the selected time frame being won by the home team, it is clear that in general, the home team is favoured to win. Figure 2 shows the home win rate for all the season included in the data. It shows that there seems to be a decline in the home win rate. This same decline was also found by Swartz and Arce (2014), in their research using data ranging back to 1979. The decline thus seems to be ongoing. Lastly, Figures 7 and 8 (Appendix) show the home win rate per season, and the home win rate per team respectively.

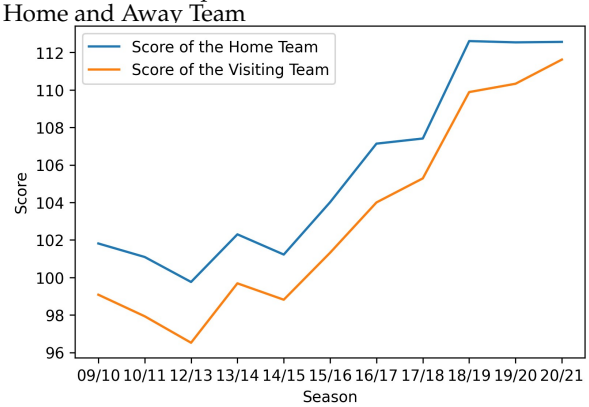
Figure 2
Home Win Rate Over the Seasons



As can be seen in Figures 2 and 7, seasons 19/20 and 20/21 are clear outliers, with a much more balanced win loss ratio for the home team. Season 20/21 could be explained by the little to no fan attendance during games (as explained in section 3.1). Season 20/21 had an average fan attendance of 1374, whereas all the other seasons average attendance was up to 17639. That's close to 13 times as many fans in attendance. Complete fan attendance distribution per team can be seen in Appendix: Figure 9. Why season 19/20 also seems to be an outlier is difficult to say. Fan attendance was on average 17789, which is basically equal to the average of the other seasons. An explanation could be that, since this season got cut short due to the pandemic, teams only played an average of 65 games during this season. It could be that home wins are more likely to happen near the end of the season, as playoff implications are bigger, and fans are more engaged. This is speculation and could be a subject for future research.

It is furthermore interesting to note that the scoring distribution for home and away teams have increased over the seasons. This can be seen in Figure 3. Figures 10, 11, and 12 (Appendix) further show that 2-point field goals have decreased, free throws have remained generally similar, whereas 3-point attempts have increased over the years. This change in shot selection is in line with the adaptation of advanced statistics. These statistics showed that it is better to avoid the 'inefficient' mid range 2-point field goal in favour of a more 'efficient' 3-point field goal. To illustrate, in the 2015-2016 NBA season, teams shot the mid range field goal at an average efficiency of 39.78%. Three point field goals were shot at an average efficiency of 36.96%[6]. Considering the mid range is worth 2 points, on average this shot attempt is worth 0.796 points, whereas, on average, a 3-point field goal attempt is worth 1.109 points. Efficiency and average field goal attempt worth, of course, changes based on players, but in general this change in perspective could explain the downward trend of two point field goals and the upwards trend of three point field goals. However, basketball analysts agree that 3-point shooting is more volatile, which could be the reason that home advantage seems to erode in this time frame. This could be a point of interest for future research.

Figure 3
Score Distribution per Season of the Home and Away Team



Lastly, the win rate of the home team is 58.49% when playing division opponents, and 58.80% when playing conference opponents. Contrastingly, win rate of the home team is 58.31% when not playing against a division opponent, and 57.58% when not playing against a conference opponent. This is already an indication that rivalry withing division or conference does not seem to affect the win rate of the home team, as speculated for research question 3.

3.4 MODEL AND FEATURE SELECTION

The predictive models used to answer the research questions are a regression decision tree and classification decision tree. The regression tree will have the score differential between the home and the visiting team as the target variable. The classification tree will have the home team 'Win' or 'Loss' as the target variable. These decision tree models were chosen for their use for practical purposes, as decision trees are very interpretable, even for people with no knowledge of machine learning or any other forms of statistic models. The visualization of these models is very clear, and even in the case that tree depth of a model is too large and visualization is no longer possible, how the decision tree model works is still easily understandable for anyone. Unknowingly or not, most people have more than likely used a decision tree in their life, or at least the conditional logic upon which a decision tree is based. This is in contrast to a more sophisticated model such as Random Forests, which go a little more in-depth into the world of machine learning and data science, which can quickly become overwhelming and more difficult to use for someone with no experience in these fields. As this research

6 Data obtained from the website of the NBA

is aimed at use for practical purposes (i.e., coaches or team managers), it is deemed best to use a more straightforward analysis model. Future research however, is very much encouraged to use more complex machine learning models to, hopefully, obtain improved results.

Secondly, this research will make use of both regression and classification models. Using multiple models broadens the scope of this research, and additionally, both the score differential used in the regression analysis, and the classification analysis of home team win or loss are used in recent research on the topic of home advantage. Whilst score differential (Loures, Shikida, and Fernandez (2021); Swartz and Arce (2014); Zimmer and Kuethe (2009)) and win/loss classification (Ehrlich and Ghimire (2020); Harris and Roebber (2019); Pollard (2002); Pollard, Prieto, and Gómez (2017)) are sometimes used without one another as ways of quantifying home advantage, most research uses both outcome variables in conjunction (e.g., Areni (2014); Demir and Rigoni (2017); Fischer and Haucap (2020); Leota et al. (2021); McSharry (2007); Ponzo and Scoppa (2018); Price and Yan (2021); Ribeiro, Mukherjee, and Zeng (2016); Van Damme and Baert (2019)).

Several feature sets will be used to train and test the models. Some of these feature sets are chosen based on literature and previous research, some are specifically selected based on the research questions, and additionally, several feature selection methods were also used to create feature sets that are supposed to be the best indicators of score differential and the home team winning. The feature sets are presented in Table 1.

Table 1
Feature sets used for analysis. Table includes the number of features included per set, textual description of each set, and the individual variables included per feature set.
For a more in-depth explanation of variables, consult Table 7 in the Appendix

Feature Set	Number of Features	Description	Variables included in the set
Basic Features (1)	26	Consists of basic box score statistics, such as: field goals, assists, turnovers and rebounds	Home FG, Home FGA, Home FG%, Home 3P, Home 3PA, Home 3P%, Home FT, Home FTA, Home FT%, Home ORB, Home TRB, Home AST, Home TOV, Opp FG, Opp FGA, Opp FG%, Opp 3P, Opp 3PA, Opp 3P%, Opp FT, Opp FTA, Opp FT%, Opp ORB, Opp TRB, Opp AST, Opp TOV
Advanced Features (2)	15	Consists of advanced box score statistics. Mostly consistent of advanced equivalents of the basic box score statistics with a few additions such as pace and FT/3PA rate	Pace, FTr, 3PAr, TRB%, AST%, STL%, BLK%, Home eFG%, Home TOV%, Home ORB%, Home FT/FGA, Opp eFG%, Opp TOV%, DRB%, Opp FT/FGA

(Continues on next page)

Feature Set	Number of Features	Description	Variables included in the set
Location Features (3)	33	Consists of the basic features (set 1), and includes all variables connected to location	Basic Features + Elevation diff, Distance diff, Home elevation, Conference, Division, Time diff, Attendance
RQ1 Features (4)	32	Consists of the basic features (set 1), and includes the basic box score variations of steals, blocks, and fouls	Basic Features + Home STL, Home BLK, Home PF, Opp STL, Opp BLK, Opp PF
RQ1 Advanced Features (5)	34	Consists of set 1 and 4, whilst adding the advanced variation of steal and block statistics. There are no advanced variation of foul statistics	Basic Features + RQ1 Features + STL%, BLK%
Top 11 Selection Score Differential (6)	11	The features selected by all feature selection methods as explained on page 14. Used in regression models on score differential	TRB%, Opp eFG%, Opp FG%, Opp FG, Opp 3P%, Home FG%, Home 3P%, Opp FT/FGA, Opp AST, Home eFG%, Home FG
Top 25 Selection Score Differential (7)	25	The features selected by three or more feature selection methods as explained on page 14. Used in regression models on score differential	Top 11 Selection ScoreDiff + Pace, Opp TRB, Opp TOV%, Opp STL, Opp FT%, Opp FT, Opp 3P, Home TRB, Home TOV%, Home FT/FGA, Home FT%, Home FT, Home AST, Home 3P
Top 10 Selection Win Loss (8)	10	The features selected by all feature selection methods as explained on page 14. Used in classification models on home win or loss	Opp eFG%, Opp FT, Opp FG%, Opp FG, Opp 3P%, Home FG%, Home FG, Opp FT/FGA, Home FT, Home 3P%
Top 19 Selection Win Loss (9)	19	The features selected by three or more feature selection methods as explained on page 14. Used in classification models on home win or loss	Top 10 Selection WL + TRB%, Opp TRB, 'Opp TOV, Opp 3P, Home eFG%, Home TRB, Home TOV%, Home AST, Home 3P

To enable the possibility to answer the main research question, set 1 was created. This set consists of all the basic box score statistics available in the data set. Statistics included

in this set are considered 'counting stats'. They include field goals, free throws, assists, rebounds, and turnovers. These statistics provide a quick breakdown of a player's or team's performance. These stats are most often cited by the casual fan when discussing players or teams. These basic box score statistics are the predominantly used statistics in research on home advantage. As such, this set will function as the baseline model for the other feature sets. Contrastingly, set 2 was created of all the advanced box score statistics. The advanced box score incorporates more advanced statistics, such as pace of the game, a more mathematical calculation of field goal efficiency, percentage of field goal attempts from 3-point range (as explained in section 3.3 on mid range vs three point shooting), and more. The difference in performance between set 1 and 2 will help answer sub research question 2.

Secondly, sets 4 and 5 were created to answer sub research question 1. Set 4 incorporates all the features from set 1, but also includes steals, blocks, and fouls of the home and away team. This set thus incorporates all statistics of the regular box score, used for NBA games. Set 5 includes these same features, but then also adds the advanced variations of statistics on steals and blocks. These two sets and the basic features set will be compared, to see if adding steals, blocks, and fouls contribute to the home team advantage. The analysis of these sets help answer the combination of sub research questions 1 and 2.

To answer sub research question 3 on the effect of location on the home advantage, feature set 3 was created. This set includes all available data on location, such as elevation, distance, conference or division rivalry, time zone differences, and attendance. These additional variables are combined with set 1, to analyse the added benefit these variables have on the home team winning advantage.

Additionally, several feature selection methods were implemented to see if combinations of statistics that literature and experts have not thought of, would better predict home team advantage. A supplemental benefit is that this will reduce the computational cost of these models. For both models, four selection methods were applied: chi-squared ('Chi-2'), logistic regression recursive wrapper based ('RFE'), Lasso ('Logistics'), and Random Forest Classifier ('Random Forest').

These methods were chosen to try several methods of feature selection. There are three primary feature selection techniques: 1) Filter methods, 2) Wrapper methods, and 3) Embedded methods. Filter based methods specify a certain metric, on which they filter features. Wrapper based methods follow a greedy search approach by evaluating all the possible combinations of features against a evaluation criterion. The best performing combination of features according to this criterion is selected. Embedded methods are considered a more hybrid method, in the sense that they combine the qualities of filter and wrapper methods. This method is then implemented by algorithms that have their own built-in feature selection method. The decision to implement multiple methods and to choose features that are chosen by a multiple of these selection methods, is to try and negate most of the negative aspects of these methods. The specific methods that were chosen (Chi-2, RFE, Lasso, Random Forest) were based on an article by Agarwal (2019). This article was also partly used for implementation of these methods in this research.

If a feature was selected as one of the top features, it would print out a 'True'. The output of the feature selection methods can be found in Tables 9 and 10 in the Appendix. From these selections, two feature sets were created for both models. The first set being the feature set where only features were selected if all selection methods returned that feature. The second set being the feature set where the features were selected by three or more of the methods. The results of these methods and the feature selection for the

regression trees can be found in Table 9. The selection for the classification trees can be found in Table 10.

From these selections, sets 6, 7, 8, and 9 were created. It is interesting to note that the top selected features are all a subset of the Basic and Advanced feature sets, with the sole exception of steals of the visiting team ('Opp_STL') in the Top 25 Selection for Score Differential feature set (set 7). This already suggests that the additional features of RQ1 and RQ3 are likely not impactful in predicting the home team advantage. Lastly, the expectation is that the models created with these sets based on feature selection should perform best, as that is of course the purpose behind feature selection.

3.5 MODEL EVALUATION

As mentioned before, the 'Basic Features' set is used as the baseline for model evaluation. Initially, multiple linear and logistic regression were conducted. However, due to high multicollinearity in the data set, these methods proved to be ineffective as regression coefficient were illogical and unreadable.

Secondly, the regression decision tree models are evaluated based on R2 score, Mean Absolute Error (MAE) and Mean Squared Error (MSE). The classification decision tree models are evaluated based on their accuracy, precision, recall, F1-scores, and ROC AUC score. These evaluation criteria are are a good fit with the data and are the most popular used metrics for both regression and classification task evaluation, which could provide the ability for comparison with future research.

Furthermore, the data was split into a train and test set, at a split of 80% and 20% respectively.

3.6 HYPERPARAMETER TUNING

A paper published by Mantovani et al. (2018) showed through an empirical study on hyperparameter tuning of decision trees that the CART algorithm (the algorithm Scikit-learn implements (Pedregosa et al. 2011)) is surprisingly sensitive to hyperparameter tuning. As such, to optimise performance of the models, hyperparameter tuning must be conducted.

The quality of a node split within the classifier decision trees are based on the Gini impurity, which measures the divergences between the probability distributions of the target values and splits a node such that it gives the least amount of impurity. Gini impurity was chosen over Information gain as, according to Raileanu and Stoffel (2004), these two metrics disagree only in 2% of all cases and the authors conclude that there is no significant difference between the two criteria. However, Gini impurity often computes quicker, as calculation of Information gain requires a logarithmic function to be computed. For the regressor decision trees, quality of node splits is based on the Mean Squared Error, one of the most commonly used metrics for splitting nodes in a regressor decision tree. This error is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node.

Using 5-fold grid search cross validation, all decision tree models were tuned for a tree depth ranging from values of 1 to 20, the minimum samples required to split an internal node ranging in values of 10 through 60 with steps of 10, and the minimum number of samples required to be at a leaf node ranging in values from 1 to 4. A split point at any depth will only be considered if it leaves at least this minimum training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression. The chosen values for these hyperparameters can

be found in Tables 2 and 3, for the regression decision tree and classification decision tree respectively. The values chosen here were chosen in part because they are largely in accordance with the findings of Mantovani et al. (2018).

4. RESULTS

Tables 2 and 3 show the evaluation scores of the regression and classification decision tree models, trained and tested on the various feature sets. Best scores on the evaluation metrics are highlighted in bold.

The optimal max depth was always found to be at least 8 or higher, which makes visualization of the decision trees impossible. To counter this, Figures 13 through 16 (Appendix) plot the first three layers of decision nodes of several models, to give an indication into the workings of the models. Furthermore, Figures 17 and 18 (Appendix), show the specific path a random sample in the data set would follow through the specified model. These figures also show what the actual score differential of the sample was for the regression trees and shows if the sample home team won or lost for the classification trees.

Table 2 shows the best performing regression models to be the models with the Advanced Features data and the Top 25 Selection Features, with R2 scores of 0.709 and 0.712, mean absolute error of 5.630 and 5.642, and mean squared error 53.257 and 52.589 respectively. The baseline of the regression model, the Basic Feature set model, performed worse with a lower R2 score of 0.642, higher mean absolute error of 6.402, and a higher mean squared error of 65.363.

Table 2
Regression Decision Tree Scoring Metrics

	Depth	Min. Leaf	Min. Split	Train R2 Score	Test R2 Score	MAE	MSE
Basic Features	18	4	30	0.855	0.642	6.402	65.363
Advanced Features	13	4	30	0.874	0.709	**5.630**	53.257
Location Features	17	4	30	0.855	0.637	6.423	66.224
RQ1 Features	14	4	30	0.857	0.636	6.460	66.466
RQ1 Advanced Features	14	4	30	0.858	0.634	6.460	66.913
Top 10 Features	10	4	40	0.810	0.692	5.915	56.301
Top 25 Features	13	4	30	**0.885**	**0.712**	5.642	**52.589**

Using a regression model, the Location, RQ1, and RQ1 Advanced feature models all perform worse than the baseline of the Basic Features model. The baseline model has a higher R2 score, lower MAE and lower MSE. The other two models do perform better than the baseline. This shows that the use of advanced statistics in analysis on the NBA should be the new status quo. Furthermore, it shows that steals, blocks, fouls, and

location features might not result in a bigger home advantage, as previous studies have speculated.

Table 3
Classification Decision Tree Scoring Metrics

	Depth	Min Samples Leaf	Min Samples Split	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score	ROC AUC Score
Basic Features	9	4	10	0.914	0.811	0.81	0.80	0.81	0.87
Advanced Features	11	4	10	0.945	**0.837**	**0.83**	**0.83**	**0.83**	0.88
Location Features	9	3	20	0.908	0.806	0.80	0.80	0.80	0.86
RQ1 Features	8	3	10	0.893	0.811	0.81	0.80	0.80	0.87
RQ1 Advanced Features	8	3	10	0.893	0.812	0.81	0.81	0.81	0.87
Top 10 Features	10	1	10	0.935	0.832	**0.83**	**0.83**	**0.83**	0.86
Top 19 Features	9	3	20	0.915	0.821	0.82	0.82	0.82	**0.89**

Table 3 shows the best performing classification model to be the model with the Advanced Features data, with an accuracy of 0.837, precision, recall, and F1-Score of 0.83 all. The Top 10 Features model performs identical in these last three metrics but has a slightly lower accuracy at 0.832. In like manner to the regression models, the Location, RQ1 and RQ1 Advanced models perform comparable to the baseline model with all evaluation metrics being nearly identical. Again, this shows that these factors likely do not impact the home advantage. Interestingly, the Advanced model performs better than the Top 10 features model for the classification task. This seems to indicate that adding more variables to the model (15 vs 10) increases performance. This is in line with the Top 25 model performing best.

Next, Tables 4 and 5 show the top 10 feature importance for all decision tree models. The most important features in the regression models (Table 4) are focused on the shooting efficiency of the home and away team, with Opp_FG%, Home_FG%, or their advanced variants of Opp_eFG% and Home_eFG% shown as the top feature importance of all the regression models. These features create the most informative split as the top two nodes for the decision trees. Secondly, rebounding seems to be an important feature for the home team, as this feature shows up near the top for most of the models. For neither the location feature sets, nor the RQ1 feature and RQ1 Advanced feature sets, do the added variables (location variables and steals/blocks/fouls) appear in the top 10 most important features for the models[7]. Elevation difference had an importance rating of only 0.0023, followed by attendance (0.0017), distance difference (0.0008), closed out by division rivalry at 0.0001. Time zone difference, conference rivalry and absolute elevation of the home team added nothing to the model, having a score of 0.0. The added importance to the model of these variables is minute.

Similarly, blocks and fouls also add little to the models. For the RQ1 Features model, visiting team fouls scored 0.0015, home team fouls scored 0.008, and blocks

7 Feature importance values outside the top 10 are not depicted in any table

of the visiting and home team scored 0.0005 and 0.0003 respectively. For the RQ1 Advanced Features model, the above-mentioned variables had a relative importance score of 0.0011, 0.0007, 0.0004, and 0.0003 respectively. Finally, the advanced statistics for home team steals (STL%) and blocks (BLK%) in the RQ1 Advanced model score 0.0092 and 0.0018.

Table 4
Top 10 Feature Importance for the Regression Decision Trees

Basic Features		**Advanced Features**		**Location Features**		**RQ1 Features**	
Opp_FG%	0.3355	Home_eFG%	0.3673	Opp_FG%	0.3353	Opp_FG%	0.3343
Home_FG%	0.3091	Opp_eFG%	0.3638	Home_FG%	0.3088	Home_FG%	0.3066
Home_FG	0.0723	TRB%	0.0766	Home_FG	0.0719	Home_FG	0.0713
Opp_FG	0.0458	Opp_TOV%	0.0747	Opp_FG	0.0456	Opp_FG	0.0455
Home_FT	0.0347	Home_TOV%	0.0729	Home_FT	0.0343	Home_FT	0.0334
Opp_FT	0.0246	Home_FT/FGA	0.0115	Opp_FT	0.0241	Opp_FT	0.0233
Opp_TRB	0.0224	Opp_FT/FGA	0.0078	Opp_TRB	0.0223	Opp_TRB	0.0233
Opp_3P	0.0193	DRB%	0.0057	Opp_3P	0.0194	Opp_3P	0.0205
Opp_TOV	0.0176	Home_ORB%	0.0057	Opp_TOV	0.0178	Home_3P	0.0164
Home_3P	0.0157	STL%	0.0039	Home_3P	0.0158	Opp_TOV	0.0160

RQ1 Advanced Features		**Top 11 Features**		**Top 25 Features**	
Opp_FG%	0.3340	Home_eFG%	0.3690	Home_eFG%	0.3350
Home_FG%	0.3068	Opp_eFG%	0.3449	Opp_eFG%	0.3143
Home_FG	0.0709	Opp_FG	0.0996	Opp_FG	0.0721
Opp_FG	0.0453	Home_FG	0.0873	Home_FG	0.0615
Home_FT	0.0331	TRB%	0.0404	TRB%	0.0394
Opp_FT	0.0223	Opp_FT/FGA	0.0343	Home_TOV%	0.0346
Opp_TRB	0.0222	Home_3P%	0.0073	Opp_TOV%	0.0344
Opp_3P	0.0199	Opp_3P%	0.0053	Home_FT	0.0263
Home_3P	0.0170	Opp_FG%	0.0041	Opp_FT	0.0202
Opp_TOV	0.0144	Home_FG%	0.0040	Home_FT/FGA	0.0091

For reference:
'Home' - home team, 'Opp' - away team, 'FG/FT/3P' - metrics related to shooting and efficiency, 'RB' - rebounds, 'TOV' - turnover, 'STL' - steals. For a more detailed explanation, consult Table 7 in the Appendix.

As such, steals seem to be the only somewhat important added feature in these models, with home team steals scoring 0.0062, and visiting team steals scoring 0.0031 in the RQ1 model, and home team steals scoring 0.0024, and visiting team steals scoring 0.0028 for the RQ1 Advanced model. The advanced steal metric (STL%) within the Advanced Features model, shows as the 10^{th} most important factor at a score of 0.0039.

Table 5
Top 10 Feature Importance for the Classification Decision Trees

Basic Features		Advanced Features		Location Features		RQ1 Features	
Opp_FG%	0.3027	Opp_eFG%	0.2942	Opp_FG%	0.3085	Opp_FG%	0.3271
Home_FG%	0.2901	Home_eFG%	0.2930	Home_FG%	0.2953	Home_FG%	0.3143
Home_FT	0.0505	TRB%	0.1008	Home_FT	0.0494	Home_FT	0.0484
Opp_FT	0.0396	Home_TOV%	0.0942	Opp_FT	0.0373	Opp_FT	0.0361
Home_TOV	0.0344	Opp_TOV%	0.0831	Opp_TOV	0.0333	Opp_TOV	0.0297
Opp_TOV	0.0337	Home_FT/FGA	0.0425	Home_TOV	0.0327	Home_TOV	0.0247
Home_3P	0.0247	Opp_FT/FGA	0.0259	Home_TRB	0.0244	Home_TRB	0.0223
Home_TRB	0.0241	FTr	0.0161	Home_3P%	0.0230	Home_3P%	0.0223
Home_3P%	0.0232	STL%	0.0130	Home_FG	0.0214	Home_3P	0.0213
Home_FG	.0232	Pace	0.0079	Home_3P	0.0206	Opp_3P%	0.0194

RQ1 Advanced Features		Top 11 Features		Top 25 Features	
Opp_FG%	0.3267	Home_FG%	0.2768	Opp_eFG%	0.2587
Home_FG%	0.3137	Opp_eFG%	0.2462	Home_FG%	0.2582
Home_FT	0.0484	Home_FT	0.0904	TRB%	0.0763
Opp_FT	0.0348	Opp_FG	0.0741	Home_TOV%	0.0732
Opp_TOV	0.0265	Home_FG	0.0732	Opp_TOV	0.0594
Home_TOV	0.0237	Opp_FG%	0.0731	Home_eFG%	0.0582
Home_TRB	0.0221	Opp_FT	0.0724	Opp_FG%	0.0575
Opp_3P%	0.0208	Home_3P%	0.0428	Home_FT	0.0438
Home_3P	0.0204	Opp_FT/FGA	0.0308	Opp_FT	0.0244
Home_3P%	0.0200	Opp_3P%	0.0202	Opp_FG	0.0163

For reference:
'Home' - home team, 'Opp' - away team, 'FG/FT/3P' - metrics related to shooting and efficiency, 'RB' - rebounds, 'TOV' - turnover, 'STL' - steals. For a more detailed explanation, consult Table 7 in the Appendix.

The same features show as the top importance features for the classification models (Table 5), with Opp_FG%, Home_FG%, or their eFG% counterparts claiming the top spots for all feature sets models. This shows that shooting efficiency is what matters most for winning basketball games. With the classification Location Feature model, the importance of some of the location variables increases, suggesting some basis for the claim that these variables are responsible for the home advantage. For this model, attendance shows an importance score of 0.0071, distance difference scores 0.0052, elevation difference 0.0037, and time difference this time is important at a score of 0.0011. Division and conference rivalry, and absolute home elevation have no added value to this model at a score of 0.

For the RQ1 Features model, visiting team steals shows a higher importance than any of the variables previously did, at 0.0109. Home team personal fouls scored 0.0076, followed by visiting team fouls, home team steals, home team blocks, and visiting team blocks, at scores of 0.0032, 0.0023, 0.0022, 0.0006 respectively. These importance scores are a little higher for these variables in the classification model than they are for the regression model.

For the RQ1 Advanced Features, the above-mentioned variables score similarly, with visiting team steals scoring 0.0104, fouls of the home team and visiting team at 0.0076 and 0.0033. Home team blocks score 0.0016. Home team steals and visiting team blocks score 0.0. Finally, the advanced statistics of steals (STL%) and blocks (BLK%) score 0.0101 and 0.0022 respectively.

Again, STL% seems to be important for the Advanced Feature model, at rank 9 and a score of 0.0130. These scores lend some support to the validation of sub research question 1.

Clearly, how teams are shooting the basketball is the best indicator for winning a game. Blocks and fouls add very little to a model's performance, with steals seeming to impact the game the most. Furthermore, location seems to be a nonfactor for game outcome. Feature importances were minimal for the models, and it could be argued that the Location model performs the worst of all the models examined.

5. DISCUSSION

This research aimed to examine the extend of, and contributors to, the home team advantage in the NBA and set out a few predictors based on literature review and newly available statistics. This study aimed to expand on current literature, whilst simultaneously filling a research gap in existing literature.

The main research question was stated as '*To what extend does home team advantage occur within the NBA*'. As seen in this research, and especially in section 3.3, home teams are significantly more likely to win a game than their visiting counterparts at 58.3% to 41.7%. To examine what factors cause this apparent disparity, three sub research question were established.

The first sub question focuses on research published by Harris and Roebber (2019) where they analysed the home team advantage using a neural network. This research finds similar results to their paper, as teams are more likely to win if they focus on shooting or impair the opposing team in theirs. Interestingly, Harris and Roebber stated that shooting percentages are of no relevance to the home advantage in their model. This is very much in contrast to the findings of this research, as shooting percentages were at the top of importance of the decision tree models. This seeming contrast could be grounds for future research.

Harris and Roebber then speculated that adding steals, blocks, and fouls to the models could possibly improve performance of these models. As such, this research asked, '*To what extend do blocks, steals, and fouls contribute to the advantage of the home team*'. The decision tree models find some basis for these variable to matter for the predictive capabilities of the model, or to the contribution of the home team advantage. Stealing the ball more often as the home team or decreasing the opportunities for the visiting team to steal the ball shows to have the most effect on the home advantage, as found through the feature importance of these variables. Fouls and blocks show a similar effect, howbeit of much lower importance. These findings conclude that the addition of blocks and fouls does have some importance for the models on predicting

home advantage, but are far from necessary to include in future research. Adding steal statistics to future research could prove to be effective.

Secondly, considering the rise in application of advanced statistics in the NBA, and sports in general, this research aimed to fill a research gap in literature by applying these statistics to the analysis of the home team advantage. Both the regression and classification decision tree model using the Advanced Features set performed best according to most, or all, evaluation metrics. It can therefore be concluded that advanced statistics should be preferred over the basic statistics, in further research and in practical use.

Finally, extant research has supplied various factors that are expected to create the home advantage. Most of these factors are speculated to be related to location and fan factors. The most prominently speculated location factors are the traveling effect for visiting teams, or the effect elevation can have on a player's body. Furthermore, sports fans are always talking about rivalry, and some research has speculated that this could be an explanation of the home advantage. This research therefore aimed to analyse this, by asking the sub question: '*To what extend does location affect the home team advantage? Location is defined as the absolute distance between teams, time zone difference between teams, and division/conference rivalry*'. The results of the model seem to indicate that these variables have very little to no predictive capacity for the home team advantage. The win rate of home teams when playing against a division opponent (58.49%) and a conference opponent (58.80%) are not notable higher than the win rate when not playing these rivals (58.31% and 57.58%). At a win rate difference of 1.22% for the (non) conference games, this is the only difference that remotely stands out. However, when conference rivalry was used in the decision tree models, it was a feature with no importance to any of the models. Furthermore, the location models performed worse on most evaluation metrics, and any of the location variables also had little to no importance as features to the model. Taking this together, it could be said that adding variables such as distance and elevation metrics, and time zone differences even made the model worse than the baseline version using basic statistics. As such, it can be concluded that these factors are no causes of the home team advantage, nor should they be used for modelling in future research. This is also in line with the results of Harris and Roebber (2019), who similarly found attendance and elevation to not be relevant for understanding home advantage.

Some limitations apply to the findings however, as the conscious decision was made to include the 2020/2021 season in the data. As mentioned, this season was played in the midst of the COVID-19 pandemic, so little to no fans were in attendance for the games. Figure 2 showed that this possibly had an effect on the home advantage, as the home win rate was drastically lower that year. The limitation, however, is that, due to time constraints, this data was simply incorporated into the full data set, and not enough time was spent to use this season as a contrasting case and to delve deeper into how home advantage factors influenced that specific season. Luckily, plenty of research has already arisen on this topic (e.g., Ehrlich and Ghimire (2020); Fischer and Haucap (2020); Higgs (2021); Loures, Shikida, and Fernandez (2021); Price and Yan (2021)). This natural experiment of these 'ghost' games during the pandemic is very much worthy of analysis for future research.

A second limitation is that the decision tree was specifically chosen for its interpretability. However, hyperparameter tuning found optimal tree depths at values greater than 3 or 4, which are often considered the limit of depth for visualization. As such, full tree visualization is impossible, which makes these models less useful in a practical sense (i.e., coaches or team managers). However, as stated in section 3.4 the logic behind decision trees can still be easily explained to practitioners. Furthermore, this apparent limitation was partly rectified by visualizing the first three nodes of some

of the decision trees (Appendix: Figures 13 through 16), but mostly by showing the path of a random sample through a tree (Appendix: Figures 17 and 18). The trade off between model performance maximization and model visualization should therefore be carefully considered with future research.

6. CONCLUSION

The aim of this research was to examine the advantage the home team possesses in the NBA, whilst trying to bring to light the factors that influence this relationship. Findings suggest that the home team is very much so advantaged at home, but in line with previous work on this subject, pinpointing an exact reason for this effect proves difficult. Proposed reasons, such as location variables, proved to be ineffective predictors of the home advantage. Further, implementing more statistics to the model, such as steals, blocks, and fouls, as proposed by Harris and Roebber (2019), had only limited effect on improving the performance of the decision tree models used in this research. Steals seemed to add the most to the home advantage, but were largely out shadowed by factors such as shooting efficiency and rebounding. Fouls were found to be of even less importance, with blocks basically being a non-factor.

Lastly, advanced statistics, pioneered in the NBA, are taking the sports world by storm, and are becoming more and more prevalent. Models in this research using these statistics performed best overall, and as such, this research suggests that future research should focus on these statistics over the more prevalently used basic statistics.

Ultimately, sport teams are encouraged to exploit the existing effect of the home advantage. The practical use of this research is not ground-breaking. Coaches and team managers should focus on the shooting performance of its players and hinder the shooting performance of the visiting team. The visiting team can be hindered through defensive schemes, but the home team can improve their own shooting performances by acquiring players that have historically shot the ball more efficiently. Furthermore, they should focus on the rebounding efforts of its players and create turnovers. Results from this research also faintly suggest that these turnovers could be created by focusing on steals.

References

Agarwal, R. 2019. The 5 feature selection algorithms every data scientist should know. *Medium, July*.

Anders, Anne and Kurt William Rotthoff. 2014. Is home-field advantage driven by the fans? evidence from across the ocean. *Applied Economics Letters*, 21(16):1165–1168.

Areni, Charles S. 2014. Home advantage, rivalry, and referee bias in representative rugby. *Sport, Business and Management: An International Journal*.

Chang, Wesley, Michael Ran, and Gary Smith. 2021. The impacts of home-court advantage in the nba.

Demir, Ender and Ugo Rigoni. 2017. You lose, i feel better: Rivalry between soccer teams and the impact of schadenfreude on stock market. *Journal of Sports Economics*, 18(1):58–76.

Draper, NR and H Smith. 1981. Applied regression analysis., 2nd edn (john wiley and sons: New york).

Ehrlich, Justin and Shankar Ghimire. 2020. Covid-19 countermeasures, major league baseball, and the home field advantage: Simulating the 2020 season using logit regression and a neural network. *F1000Research*, 9(414):414.

Feldman, Dan. 2018. Rumor: Spurs won't trade kawhi leonard to western conference team. https://nba.nbcsports.com/2018/06/18/rumor-spurs-wont-trade-kawhi-leonard-to-western-conference-team/.

Fischer, Kai and Justus Haucap. 2020. Does crowd support drive the home advantage in professional soccer? evidence from german ghost games during the covid-19 pandemic.

Friedman, J. 2007. Rocket science: Daryl morey brings hard-core statistical analysis to the nba. https://www.houstonpress.com/news/rocket-science-daryl-morey-brings-hard-core-statistical-analysis-to-the-nba-6540549?showFullText=true.

Gobikas, Mindaugas, Alexandru Radu, and Jonas Miklovas. 2020. Home court advantage in basketball-a case study of zalgiris kaunas basketball team.

Goldman, Matt and Justin M Rao. 2012. Effort vs. concentration: the asymmetric impact of pressure on nba performance. In *Proceedings of the MIT Sloan sports analytics conference*, pages 1–10, Citeseer.

Goumas, Chris. 2014. Home advantage in australian soccer. *Journal of Science and Medicine in Sport*, 17(1):119–123.

Greer, Donald L. 1983. Spectator booing and the home advantage: A study of social influence in the basketball arena. *Social psychology quarterly*, pages 252–261.

Harris, Austin R and Paul J Roebber. 2019. Nba team home advantage: Identifying key factors using an artificial neural network. *PloS one*, 14(7):e0220630.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Higgs, Nico. 2021. *Home Advantage in North American Professional Sports Before and During COVID-19: A Bayesian Perspective*. Ph.D. thesis, University of Saskatchewan.

Hunter, J. D. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Karney, Charles FF. 2013. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55.

Leota, Josh, Daniel Hoffman, Luis Mascaro, Mark É Czeisler, Kyle Nash, Sean Drummond, Clare Anderson, Shantha MW Rajaratnam, and Elise Facer-Childs. 2021. Home is where the hustle is: The influence of crowds on effort and home advantage in the national basketball association. *Available at SSRN 3898283*.

Lopez, Michael J, Gregory J Matthews, and Benjamin S Baumer. 2018. How often does the best team win? a unified approach to understanding randomness in north american sport. *The Annals of Applied Statistics*, 12(4):2483–2516.

Loures, Alexandre, Claudio D Shikida, and Rodrigo Fernandez. 2021. Rivalry with and without crowds: An analysis of four regional soccer tournaments in brazil before and during the covid-19 pandemic. *Available at SSRN 3822740*.

Mantovani, Rafael Gomes, Tomáš Horváth, Ricardo Cerri, Sylvio Barbon Junior, Joaquin Vanschoren, and André Carlos Ponce de Leon Ferreira de Carvalho. 2018. An empirical study

on hyperparameter tuning of decision trees. *arXiv preprint arXiv:1812.02207*.
McHill, Andrew W and Evan D Chinoy. 2020. Utilizing the national basketball association's covid-19 restart "bubble" to uncover the impact of travel and circadian disruption on athletic performance. *Scientific Reports*, 10(1):1–7.
McKinney, Wes et al. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, Austin, TX.
McSharry, Patrick E. 2007. Effect of altitude on physiological performance: a statistical analysis using results of international football games. *Bmj*, 335(7633):1278–1281.
Neave, Nick and Sandy Wolfson. 2003. Testosterone, territoriality, and the 'home advantage'. *Physiology & behavior*, 78(2):269–275.
Nevill, Alan M, Nigel J Balmer, and A Mark Williams. 2002. The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of sport and exercise*, 3(4):261–272.
Oberhofer, Harald, Tassilo Philippovich, and Hannes Winner. 2010. Distance matters in away games: Evidence from the german football league. *Journal of Economic Psychology*, 31(2):200–211.
Paine, N. 2013. Where utah, denver lose an edge. https://www.espn.com/nba/insider/story/_/id/9014283/nba-analyzing-real-home-court-advantage-utah-jazz-denver-nuggets.
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
Pollard, Richard. 2002. Evidence of a reduced home advantage when a team moves to a new stadium. *Journal of Sports Sciences*, 20(12):969–973.
Pollard, Richard and Vasilis Armatas. 2017. Factors affecting home advantage in football world cup qualification. *International Journal of Performance Analysis in Sport*, 17(1-2):121–135.
Pollard, Richard, Jaime Prieto, and Miguel-Ángel Gómez. 2017. Global differences in home advantage by country, sport and sex. *International Journal of Performance Analysis in Sport*, 17(4):586–599.
Ponzo, Michela and Vincenzo Scoppa. 2018. Does the home advantage depend on crowd support? evidence from same-stadium derbies. *Journal of Sports Economics*, 19(4):562–582.
Price, Michael and Jun Yan. 2021. The effects of the nba covid bubble on the nba playoffs: A case study for home-court advantage. *arXiv preprint arXiv:2103.02832*.
Raileanu, Laura Elena and Kilian Stoffel. 2004. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93.
Ribeiro, Haroldo V, Satyam Mukherjee, and Xiao Han T Zeng. 2016. The advantage of playing home in nba: Microscopic, team-specific and evolving features. *PloS one*, 11(3):e0152440.
Roeder, O. 2017. Do nba refs favor the home team? https://pudding.cool/2017/03/home-court/.
Schwartz, Barry and Stephen F Barsky. 1977. The home advantage. *Social forces*, 55(3):641–661.
Seabold, Skipper and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.
Snyder, Eldon E and Dean A Purdy. 1985. The home advantage in collegiate basketball. *Sociology of sport journal*, 2(4):352–356.
Swartz, Tim B and Adriano Arce. 2014. New insights involving the home team advantage. *International Journal of Sports Science & Coaching*, 9(4):681–692.
Van Damme, Nils and Stijn Baert. 2019. Home advantage in european international soccer: Which dimension of distance matters? *Economics*, 13(1).
Waskom, Michael, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. 2017. mwaskom/seaborn: v0.8.1 (september 2017).
Willoughby, Jack and C Becker. 2014. Attendance, home advantage, and the effect of a city on its professional sports teams. *Duke University Durham, North Carolina*.
Winfree*, Jason A, Jill J McCluskey, Ron C Mittelhammer, and Rodney Fort. 2004. Location and attendance in major league baseball. *Applied Economics*, 36(19):2117–2124.

Zimmer, Timothy and Todd H Kuethe. 2009. Testing for bias and manipulation in the national basketball association playoffs. *Journal of Quantitative Analysis in Sports*, 5(3).

APPENDIX

Table 6
NBA Team Abbreviation

Full team name	Abbreviation
Atlanta Hawks	ATL
Boston Celtics	BOS
Brooklyn Nets	BRK
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	POR
New York Knicks	NYK
Oklahoma City Thunder	OKC
Orlando Magic	ORL
Philadelphia 76ers	PHI
Phoenix Suns	PHX
Portland Trailblazers	POR
Sacramento Kings	SAC
San Antonio Spurs	SAS
Toronto Raptors	TOR
Utah Jazz	UTA
Washington Wizards	WAS

Table 7
Variables used in the data set and associated variable explanation

Variable	Explanation
Home	The name of the home team. Abbreviated according to Table 1.
Opp	The name of the visiting team. Abbreviated according to Table 1.
W/L	Categorical indicator whether the home team won or lost the game.
Score_Home	The final score of the home team.
Score_Opp	The final score of the visiting team.
Score_diff	The final score differential between the teams.
Home_FG	Made field goals by the home team.
Home_FGA	Attempted field goals by the home team.
Home_FG%	Field goal percentage of the home team. Calculated as FG / FGA.
Home_3P	Three point field goals made by the home team.
Home_3PA	Three point field goals attempted by the home team.
Home_3P%	Three point field goals percentage by the home team. Calculated as 3P / 3PA.
Home_FT	Free throws made by the home team.
Home_FTA	Free throws attempted by the home team.
Home_FT%	Free throw percentage by the home team. Calculated as FT / FTA.
Home_ORB	Offensive rebounds by the home team.
Home_TRB	Total rebounds by the home team. Includes offensive and defensive rebounds.
Home_AST	Total assists by the home team.
Home_STL	Total steals by the home team.
Home_BLK	Total blocks by the home team.
Home_TOV	Total turnovers by the home team.
Home_PF	Total personal fouls by the home team.
Opp_FG	Made field goals by the visiting team.
Opp_FGA	Attempted field goals by the visiting team.
Opp_FG%	Field goal percentage of the visiting team. Calculated as FG / FGA.
Opp_3P	Three point field goals made by the visiting team.
Opp_3PA	Three point field goals attempted by the visiting team.
Opp_3P%	Three point field goals percentage by the visiting team. Calculated as 3P / 3PA.
Opp_FT	Free throws made by the visiting team.
Opp_FTA	Free throws attempted by the visiting team.
Opp_FT%	Free throw percentage by the visiting team. Calculated as FT / FTA.
Opp_ORB	Offensive rebounds by the visiting team.
Opp_TRB	Total rebounds by the visiting team. Includes offensive and defensive rebounds.
Opp_AST	Total assists by the visiting team.
Opp_STL	Total steals by the visiting team.
Opp_BLK	Total blocks by the visiting team.

(Continues on next page)

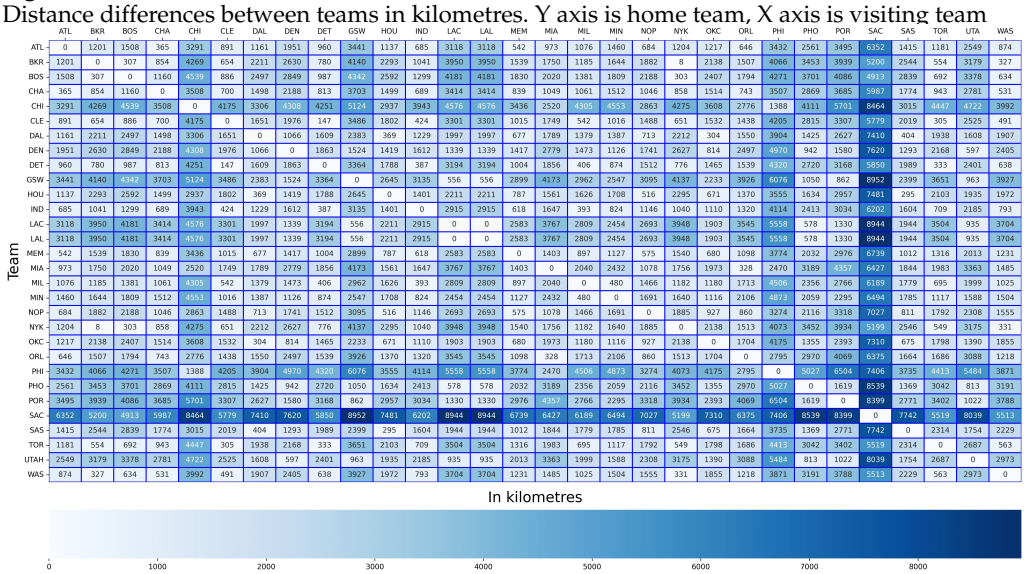
Table 7 cntd. Variables used in the data set and associated variable explanation

Variable	Explanation
Opp_TOV	Total turnovers by the visiting team.
Opp_PF	Total personal fouls by the visiting team.
Pace	An estimate of possessions per 48 minutes by the home team.
FTr	Number of FT Attempts Per FG Attempt by the home team.
3PAr	Percentage of FG Attempts from 3-Point Range by the home team.
TS%	A measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws by the home team.
TRB%	An estimate of the percentage of total available rebounds the home team grabbed.
AST%	An estimate of the percentage of field goals the home team assisted on.
STL%	An estimate of the percentage of opponent possessions that end with a steal by the home team.
BLK%	An estimate of the percentage of opponent two-point field goal attempts were blocked by the home team.
Home_eFG%	Effective Field Goal Percentage of the home team. This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.
Home_TOV%	Turnover Percentage of the home team. An estimate of turnovers committed per 100 plays.
Home_ORB%	Offensive Rebound Percentage of the home team. An estimate of the percentage of available offensive rebounds the home team grabbed.
DRB%	Defensive Rebound Percentage. An estimate of the percentage of available defensive rebounds the home team grabbed.
Home_FT/FGA	Free Throws Per Field Goal Attempt of the home team.
Opp_eFG%	Effective Field Goal Percentage of the visiting team.
Opp_TOV%	Turnover Percentage of the visiting team.
Opp_FT/FGA	Free Throws Per Field Goal Attempt of the visiting team.
Attendance	Total fan attendance for that game.
Elevation_diff	Difference in elevation between the arenas of the home and visiting team. In kilometres.
Distance_diff	Difference in distance between the arenas of the home and visiting team. In kilometres.
Home_elevation	Elevation of the arena of the home team. In kilometres.
Conference	Categorical dummy variable to indicate whether the home and visiting are in the same conference.
Division	Categorical dummy variable to indicate whether the home and visiting are in the same division.
Time_diff	Difference in time zone between the home and visiting team.
Season	Indicator to show in what season the game was played. Used for visualization purposes.

Table 8
Descriptive statistics of the data set

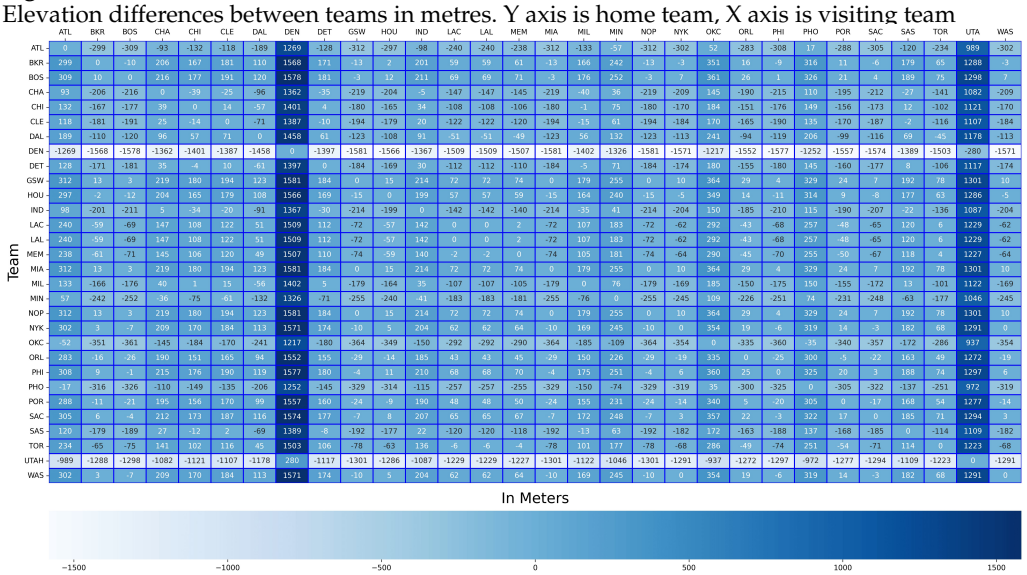
	count	mean	std	min	25%	50%	75%	max
Score_Home	13100.0	105.46	12.94	59.00	96.00	105.00	114.00	161.00
Score_Opp	13100.0	102.89	12.94	56.00	94.00	103.00	111.00	168.00
score_diff	13100.0	2.57	13.66	-57.00	-7.00	4.00	11.00	61.00
Home_FG	13100.0	39.21	5.25	19.00	36.00	39.00	43.00	63.00
Home_FGA	13100.0	84.75	7.60	59.00	80.00	85.00	90.00	125.00
Home_FG%	13100.0	0.46	0.06	0.27	0.42	0.46	0.50	0.68
Home_3P	13100.0	9.12	4.06	0.00	6.00	9.00	12.00	28.00
Home_3PA	13100.0	25.22	8.67	4.00	19.00	25.00	31.00	70.00
Home_3P%	13100.0	0.36	0.10	0.00	0.29	0.36	0.42	0.89
Home_FT	13100.0	17.93	6.10	1.00	14.00	17.00	22.00	45.00
Home_FTA	13100.0	23.49	7.50	1.00	18.00	23.00	28.00	64.00
Home_FT%	13100.0	0.76	0.10	0.14	0.70	0.77	0.83	1.00
Home_ORB	13100.0	10.62	3.85	1.00	8.00	10.00	13.00	27.00
Home_TRB	13100.0	43.83	6.62	17.00	39.00	44.00	48.00	72.00
Home_AST	13100.0	23.34	5.17	6.00	20.00	23.00	27.00	50.00
Home_STL	13100.0	7.63	2.94	0.00	6.00	7.00	9.00	22.00
Home_BLK	13100.0	5.08	2.60	0.00	3.00	5.00	7.00	20.00
Home_TOV	13100.0	13.57	3.82	1.00	11.00	13.00	16.00	29.00
Home_PF	13100.0	19.98	4.23	6.00	17.00	20.00	23.00	41.00
Opp_FG	13100.0	38.36	5.18	20.00	35.00	38.00	42.00	60.00
Opp_FGA	13100.0	84.82	7.64	59.00	80.00	85.00	90.00	129.00
Opp_FG%	13100.0	0.45	0.05	0.27	0.42	0.45	0.49	0.67
Opp_3P	13100.0	8.95	4.02	0.00	6.00	9.00	12.00	29.00
Opp_3PA	13100.0	25.28	8.67	3.00	19.00	25.00	31.00	63.00
Opp_3P%	13100.0	0.35	0.10	0.00	0.29	0.35	0.42	0.78
Opp_FT	13100.0	17.23	5.96	1.00	13.00	17.00	21.00	52.00
Opp_FTA	13100.0	22.62	7.28	1.00	17.00	22.00	27.00	64.00
Opp_FT%	13100.0	0.76	0.10	0.18	0.70	0.77	0.83	1.00
Opp_ORB	13100.0	10.40	3.80	0.00	8.00	10.00	13.00	38.00
Opp_TRB	13100.0	42.71	6.49	20.00	38.00	43.00	47.00	81.00
Opp_AST	13100.0	22.16	5.11	4.00	19.00	22.00	26.00	46.00
Opp_STL	13100.0	7.62	2.89	0.00	6.00	7.00	9.00	20.00
Opp_BLK	13100.0	4.67	2.46	0.00	3.00	4.00	6.00	17.00
Opp_TOV	13100.0	13.76	3.87	2.00	11.00	14.00	16.00	29.00
Opp_PF	13100.0	20.60	4.35	5.00	18.00	20.00	23.00	42.00
ORtg	13100.0	110.17	11.35	64.60	102.50	110.00	117.70	155.60
DRtg	13100.0	107.51	11.47	64.10	99.70	107.50	115.30	153.50
Pace	13100.0	95.04	5.77	76.10	91.00	94.90	98.90	118.20
FTr	13100.0	0.28	0.10	0.01	0.21	0.27	0.34	0.88
3PAr	13100.0	0.30	0.09	0.04	0.23	0.29	0.36	0.68
TS%	13100.0	0.56	0.06	0.33	0.51	0.55	0.60	0.78
TRB%	13100.0	50.63	5.25	29.30	47.10	50.60	54.10	72.10
AST%	13100.0	59.41	9.98	22.60	52.80	59.50	66.70	93.30
STL%	13100.0	7.96	2.99	0.00	5.80	7.80	9.90	21.70
BLK%	13100.0	8.53	4.22	0.00	5.50	8.10	11.10	38.50
Home_eFG%	13100.0	0.52	0.07	0.28	0.47	0.51	0.56	0.77
Home_TOV%	13100.0	12.50	3.40	1.00	10.10	12.40	14.70	26.80
Home_ORB%	13100.0	24.54	7.60	2.00	19.20	24.35	29.50	55.60
Home_FT/FGA	13100.0	0.21	0.08	0.01	0.16	0.21	0.26	0.67
Opp_eFG%	13100.0	0.51	0.07	0.28	0.46	0.51	0.55	0.78
Opp_TOV%	13100.0	12.68	3.46	1.90	10.20	12.50	15.00	27.80
Opp_DRB%	13100.0	76.30	7.46	45.90	71.40	76.60	81.60	100.00
Opp_FT/FGA	13100.0	0.21	0.08	0.01	0.15	0.20	0.25	0.65
Attendance	13100.0	16,297.65	5,048.40	0.00	15,424.75	18,055.00	19,316.50	23,152.00
elevation_diff	13100.0	-0.07	500.79	-1,581.00	-147.01	0.00	147.01	1,581.00
distance_diff	13100.0	2,350.25	1,808.93	0.00	1,039.66	1,802.10	3,274.33	8,951.61
home_elevation	13100.0	202.34	350.49	1.01	11.00	79.00	215.01	1,582.00
conference	13100.0	0.63	0.48	0.00	0.00	1.00	1.00	1.00
division	13100.0	0.19	0.39	0.00	0.00	0.00	0.00	1.00
time_diff	13100.0	0.00	1.43	-3.00	-1.00	0.00	1.00	3.00

Figure 4
Distance differences between teams in kilometres. Y axis is home team, X axis is visiting team



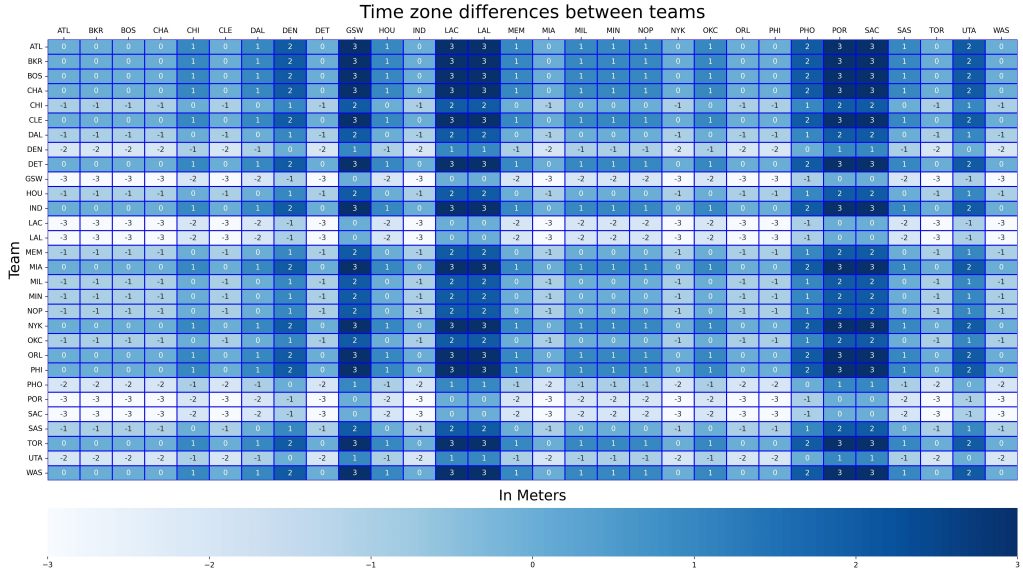
Distance differences between the NBA teams. As measured by the geodesic distance (shortest distance between two points on a curved surface, such as the Earth) between the latitudinal and longitudinal values of the NBA arenas.

Figure 5
Elevation differences between teams in metres. Y axis is home team, X axis is visiting team



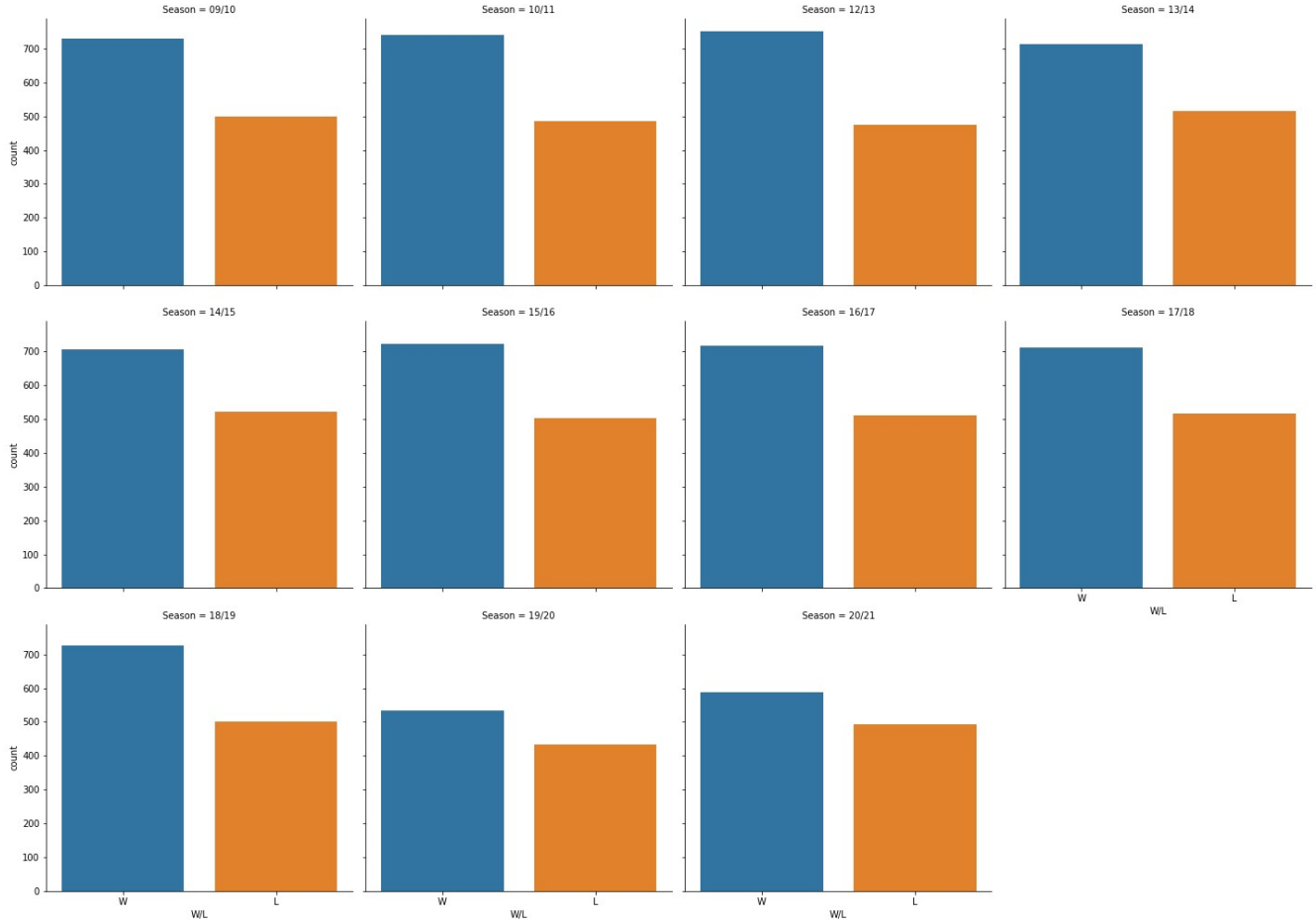
A positive value in this graph means the visiting team experiences an increase of elevation compared to the elevation of their home arena, whereas a negative values means the visiting team has a home arena that is situated at a higher altitude than the arena of the team they are visiting. As is clearly visible in this figure, Denver deservedly earned the nickname of 'Mile-High City', with the Utah Jazz, situated in Salt Lake City, being located at the second highest altitude by a large margin.

Figure 6
Time zone differences between teams. Y axis is home team, X axis is visiting team



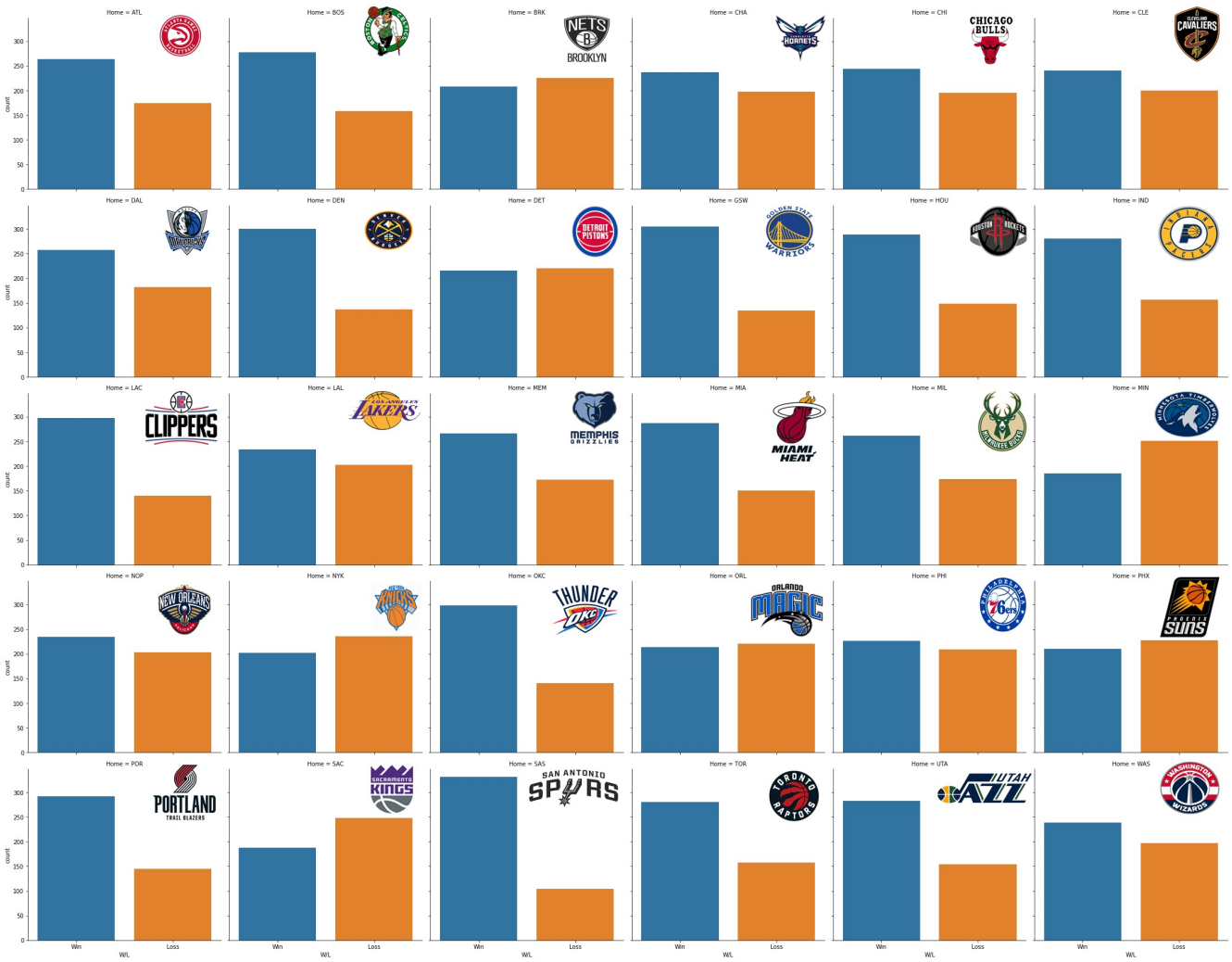
The values in this table can take on positive and negative values. A positive value means the away team travels from West to East. A negative value means the away team travels from East to West. Vertical teams are the home teams, horizontal teams are the away teams. The importance of traveling from East to West or vice versa was not analysed in this research. However, this data can be used for future research, as McHill and Chinoy (2020) alluded to.

Figure 7
Distribution of Home Win Percentage Per Season



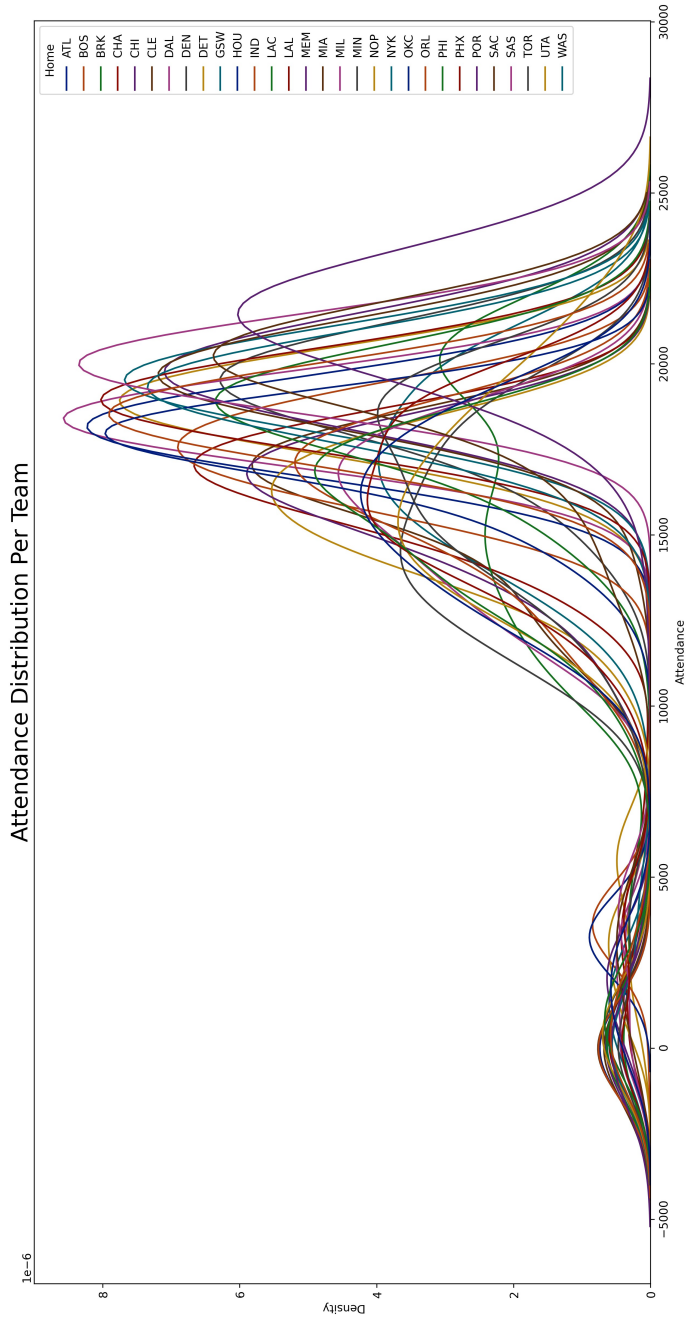
The left bar in each subplot represents a win for the home team, with the right bar representing a loss for the home team. Visible here is the evident home advantage over the seasons in the NBA. Furthermore, as mentioned in section 3.3, the last two season are clear outliers. Figure 2 shows a clearer view of the home win percentage over time.

Figure 8
Home Winning Percentage per Team



The home winning percentage per team in the NBA. Teams are listed row wise from top left to bottom right, in accordance with the order of teams in Table 6. Thus, Atlanta Hawks are in the top left subplot, the Cleveland Cavaliers top right, and the Washington Wizards bottom right.

Figure 9
Attendance Distribution per NBA Team



Attendance distribution per team over the seasons. Interesting to note here, is the distribution of attendance around the 0 and 4000 mark. These correspond to the 2020-2021 season, which was blemished by COVID-19. As pointed out in section 3.3, average attendance is 17639. Displayed here is that the teams hover around this value, with some exceptions such as the Chicago Bulls. Located in the United Center in Chicago, they have the highest capacity for fans.

Figure 10
Three Point Attempts over the Seasons

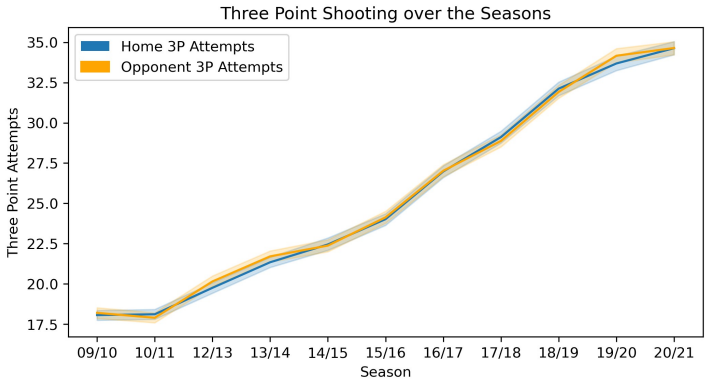


Figure 11
Two Point Attempts over the Seasons

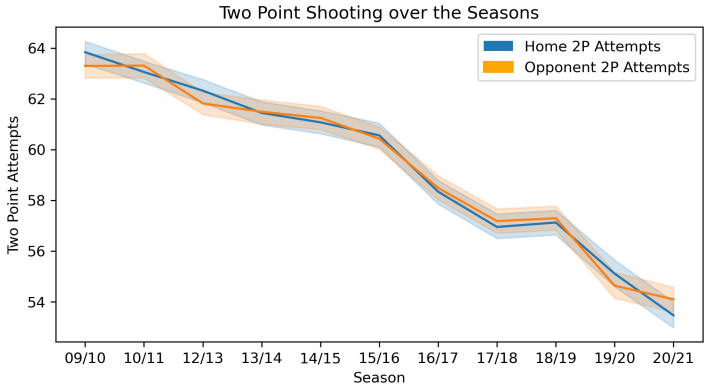


Figure 12
Free Throw Attempts over the Season

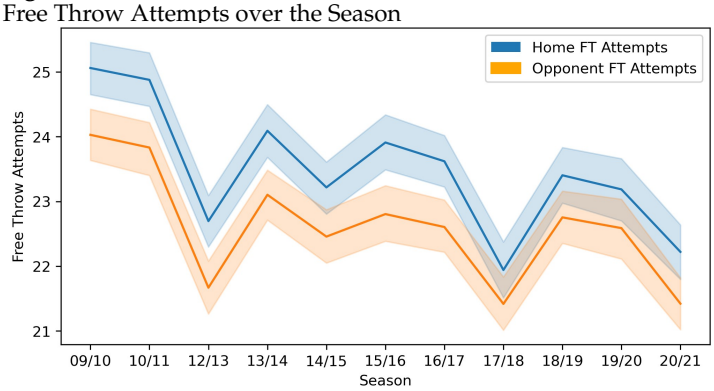


Table 9
Feature Selection for Score Differential – Regression Tree. Explanations of feature abbreviations are found in Table 7

	Feature	Chi-2	RFE	Logistics	Random Forest	Total
1	TRB%	True	True	True	True	4
2	Opp_eFG%	True	True	True	True	4
3	Opp_FT/FGA	True	True	True	True	4
4	Opp_FG%	True	True	True	True	4
5	Opp_FG	True	True	True	True	4
6	Opp_AST	True	True	True	True	4
7	Opp_3P%	True	True	True	True	4
8	Home_eFG%	True	True	True	True	4
9	Home_FG%	True	True	True	True	4
10	Home_FG	True	True	True	True	4
11	Home_3P%	True	True	True	True	4
12	Pace	False	True	True	True	3
13	Opp_TRB	True	True	True	False	3
14	Opp_TOV%	False	True	True	True	3
15	Opp_STL	True	True	True	False	3
16	Opp_FT%	False	True	True	True	3
17	Opp_FT	True	True	True	False	3
18	Opp_3P	True	True	True	False	3
19	Home_TRB	True	True	True	False	3
20	Home_TOV%	False	True	True	True	3
21	Home_FT/FGA	True	True	False	True	3
22	Home_FT%	False	True	True	True	3
23	Home_FT	True	True	True	False	3
24	Home_AST	True	True	True	False	3
25	Home_3P	True	True	True	False	3
26	STL%	True	False	False	True	2
27	Opp_TOV	False	True	True	False	2
28	Opp_PF	False	True	True	False	2
29	Opp_3PA	False	False	True	True	2
30	Home_TOV	False	True	True	False	2
31	Home_STL	True	True	False	False	2
32	Home_PF	False	True	True	False	2
33	FTr	True	False	False	True	2
34	BLK%	True	False	False	True	2
35	Attendance	False	False	True	True	2
36	home_elevation	True	False	False	False	1
37	elevation_diff	False	False	False	True	1
38	division	True	False	False	False	1
39	distance_diff	False	False	False	True	1
40	conference	True	False	False	False	1
41	Opp_DRB%	False	False	False	True	1
42	Opp_BLK	True	False	False	False	1
43	Home_ORB%	False	False	False	True	1
44	Home_FTA	True	False	False	False	1
45	Home_BLK	True	False	False	False	1
46	AST%	False	False	False	True	1
47	3PAr	False	False	False	True	1
48	time_diff	False	False	False	False	0
49	Opp_ORB	False	False	False	False	0
50	Opp_FTA	False	False	False	False	0
51	Opp_FGA	False	False	False	False	0
52	Home_ORB	False	False	False	False	0
53	Home_FGA	False	False	False	False	0
54	Home_3PA	False	False	False	False	0

Table 10
Feature Selection for Win Loss – Classification Tree. Explanations of feature abbreviations are found in Table 7

	Feature	Chi-2	RFE	Logistics	Random Forest	Total
1	Opp_eFG%	True	True	True	True	4
2	Opp_FT/FGA	True	True	True	True	4
3	Opp_FT	True	True	True	True	4
4	Opp_FG%	True	True	True	True	4
5	Opp_FG	True	True	True	True	4
6	Opp_3P%	True	True	True	True	4
7	Home_FT	True	True	True	True	4
8	Home_FG%	True	True	True	True	4
9	Home_FG	True	True	True	True	4
10	Home_3P%	True	True	True	True	4
11	TRB%	True	True	False	True	3
12	Opp_TRB	True	True	True	False	3
13	Opp_TOV	True	True	True	False	3
14	Opp_3P	True	True	True	False	3
15	Home_eFG%	True	True	False	True	3
16	Home_TRB	True	True	False	True	3
17	Home_TOV%	False	True	True	True	3
18	Home_AST	True	True	True	False	3
19	Home_3P	True	True	True	False	3
20	Opp_STL	True	False	True	False	2
21	Opp_PF	False	True	True	False	2
22	Opp_FT%	False	True	True	False	2
23	Opp_BLK	True	False	True	False	2
24	Opp_AST	True	False	True	False	2
25	Home_PF	False	True	True	False	2
26	Home_ORB%	True	False	True	False	2
27	Home_FTA	True	True	False	False	2
28	Home_FT/FGA	True	True	False	False	2
29	time_diff	False	False	True	False	1
30	home_elevation	True	False	False	False	1
31	elevation_diff	False	False	True	False	1
32	division	False	False	True	False	1
33	conference	False	False	True	False	1
34	STL%	True	False	False	False	1
35	Opp_TOV%	False	True	False	False	1
36	Opp_FTA	False	True	False	False	1
37	Opp_FGA	False	True	False	False	1
38	Home_TOV	False	True	False	False	1
39	Home_STL	True	False	False	False	1
40	Home_FT%	False	True	False	False	1
41	Home_FGA	False	True	False	False	1
42	Home_BLK	True	False	False	False	1
43	FTr	True	False	False	False	1
44	BLK%	True	False	False	False	1
45	Attendance	False	False	True	False	1
46	distance_diff	False	False	False	False	0
47	Pace	False	False	False	False	0
48	Opp_ORB	False	False	False	False	0
49	Opp_DRB%	False	False	False	False	0
50	Opp_3PA	False	False	False	False	0
51	Home_ORB	False	False	False	False	0
52	Home_3PA	False	False	False	False	0
53	AST%	False	False	False	False	0
54	3PAr	False	False	False	False	0

Figures 13 and 14 show the first three node splits of the regression decision tree models using the Advanced Feature set and the Top 25 Features Selection set (consult Table 1 for explanation of the feature sets) as examples. As mentioned in the results (section 4), and specifically Tables 4 and 5, shooting efficiency of both the home and away team are the most important for node splitting in the models. As seen in the first node, it is important to prevent your opponent from shooting well. Secondarily, the second node shows the importance of ensuring efficient/good scoring of the home team.

Figure 13
Regression Decision Tree Advanced Features. First three nodes

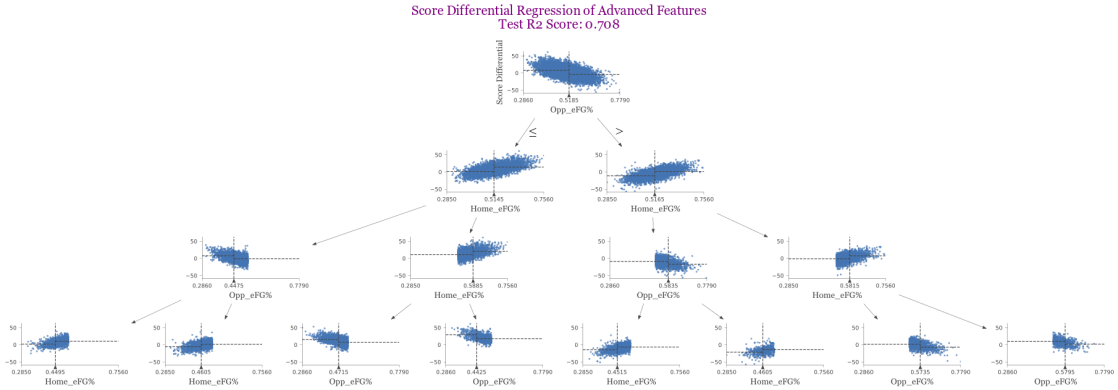
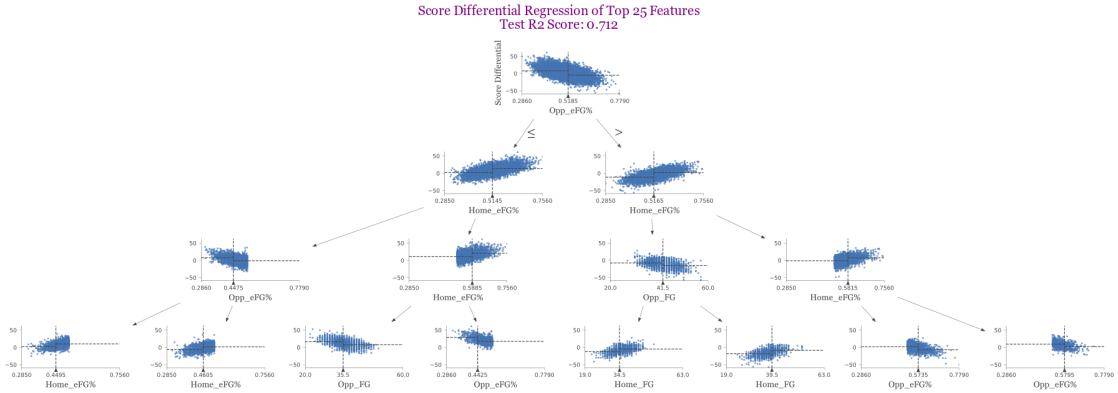


Figure 14
Regression Decision Tree Top 25 Features. First three nodes



Figures 15 and 16 show the first three node splits of the classification decision tree models using the Advanced Feature set and the Top 19 Features Selection set (consult Table 1 for explanation of the feature sets) as examples. Again, the top two nodes show the importance of limiting the shooting efficiency of the visiting team, whilst promoting scoring efficiency of the home team. Seen here also, is that the advanced statistic of TRB% (the percentage of available rebounds the home team secured) is important for the home advantage. Securing less than 53.85% of the available rebounds, increases the likelihood of a loss for the home team.

The clarity of these visualizations can be especially useful for practical uses.

Figure 15
Classification Decision Tree Advanced Features. First three nodes

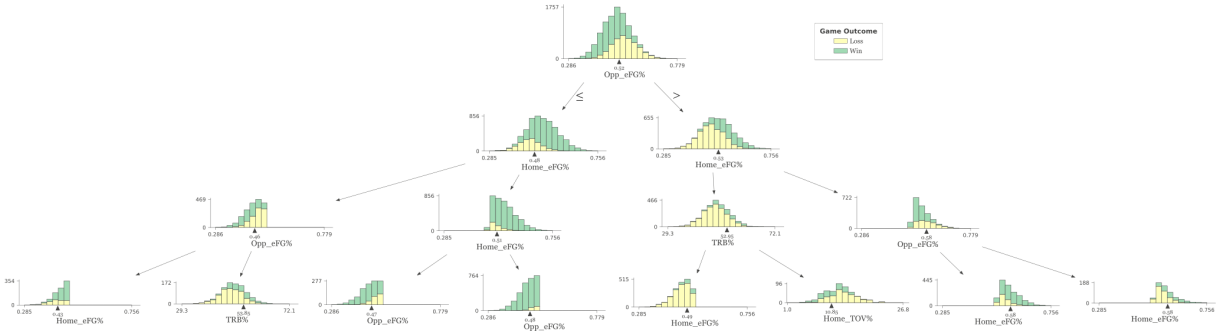


Figure 16
Classification Decision Tree Top 19 Features. First three nodes

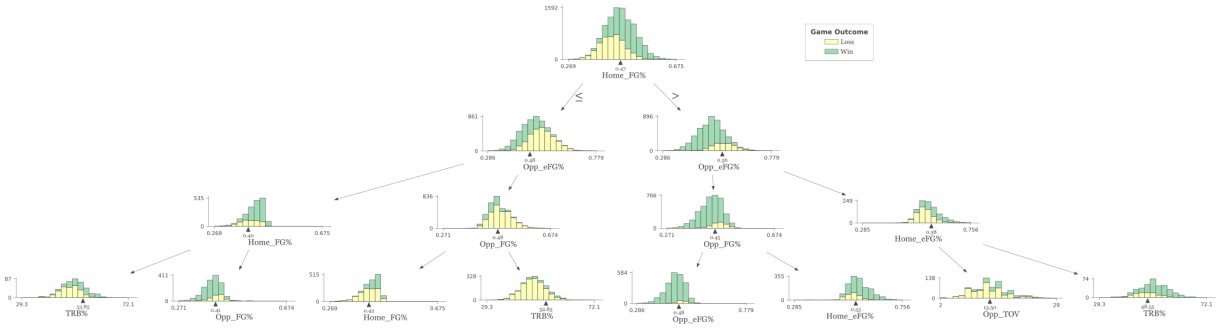
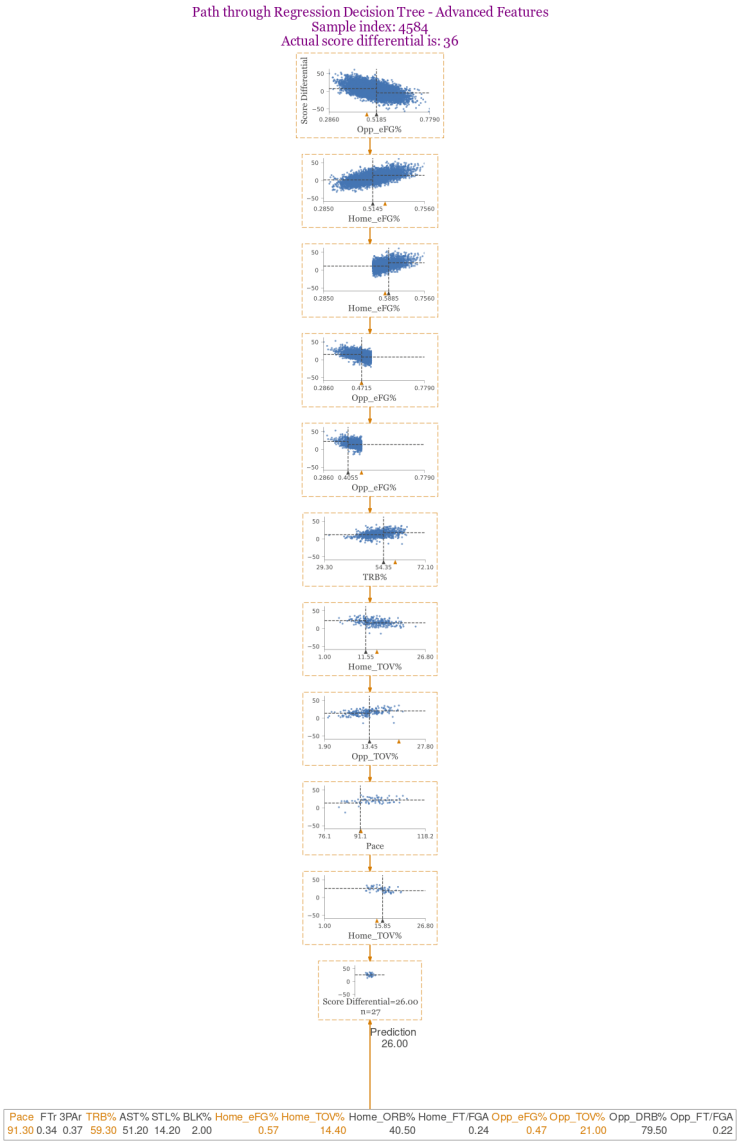
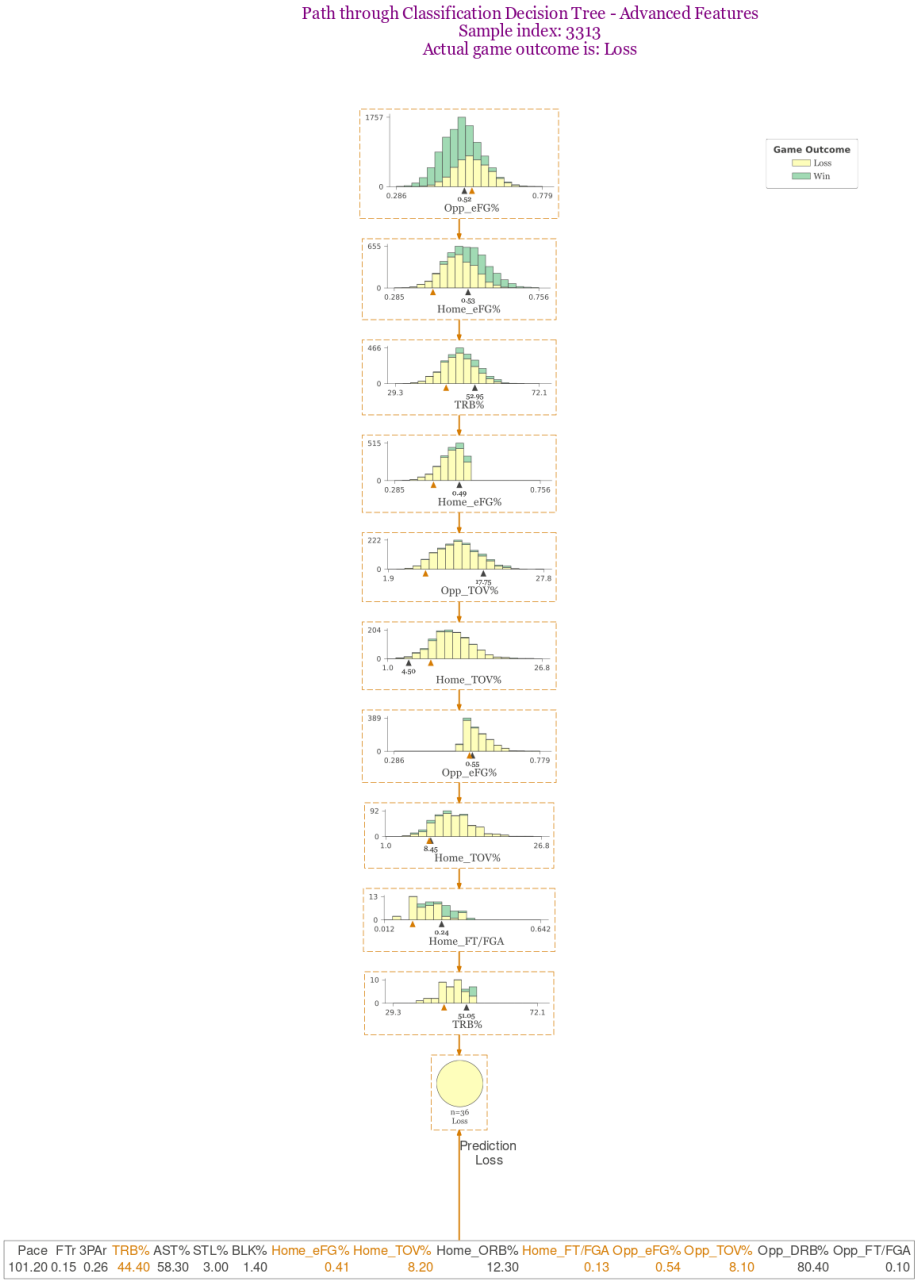


Figure 17
Path through Advanced Features Regression Tree



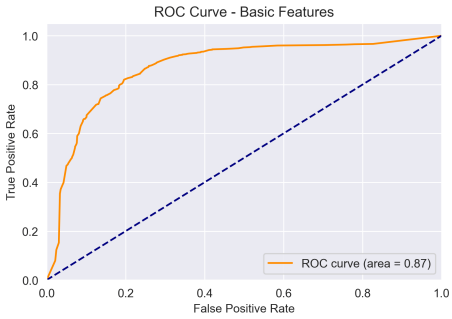
This figure shows a path through the regression decision tree that is trained on the Advanced Features feature set (Table 1). Considering the depth chosen during hyper-parameter tuning (section 3.6) does not easily allow for full tree depiction, the path of a randomly chosen sample/game is illustrated to show the logic of the model. The first few nodes show that the visiting team shot the basketball at a low efficiency ($0.47 < 0.5185$, the value chosen for the split), whilst the home team shot at high efficiency ($0.57 > 0.5145$). The path further shows that even though the home team turned the ball over more than preferred (Home_TOV% $14.40 > 11.55$), the home team made sure the visiting team turned the ball over even more (Opp_TOV% $21.00 > 13.45$). Final prediction of the model is a score differential of 26 in the favour of the home team, whilst the actual outcome of the game was a differential of 36. This specific sample is not that accurate, likely due to the fact that a score differential of 36 is quite the outlier when we compare it to the score differential mean of 2.57 (section 3.3 and Table 8).

Figure 18
Path through Advanced Features Classification Tree

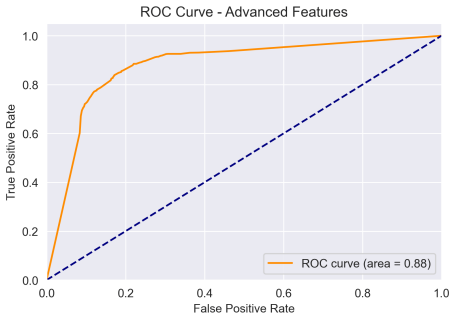


In accordance to the reasoning mentioned at Figure 17, this figure depicts a path through the classification decision tree that is trained on the Advanced Features feature set Table 1). This specific sample shows the correct classification by the model: a loss for the home team. The figure shows that the home team allowed the visiting team to shoot a little too efficient (Opp_eFG% 0.54 > 0.52), whilst simultaneously not scoring in an efficient enough manner themselves (Home_eFG% 0.41 < 0.53). Furthermore, the home team struggled a lot with securing rebounds (TRB% 44.40 < 52.95) and did not manage to make the visiting team turn the ball over (Opp_TOV% 8.10 < 17.75).

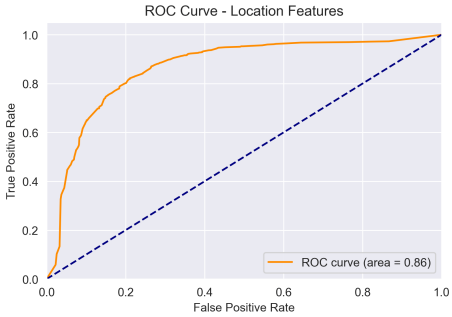
Figure 19
Receiver Operating Characteristic curves of the Classification Decision Trees



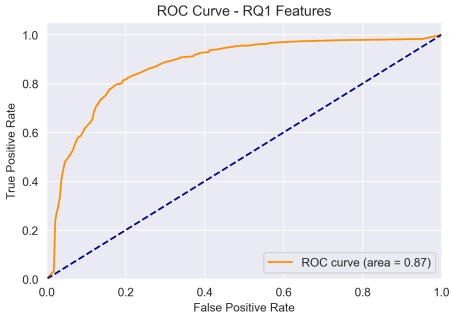
(a) Basic Features - Area = 0.87



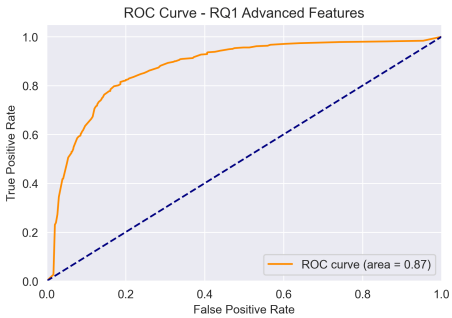
(b) Advanced Features - Area = 0.88



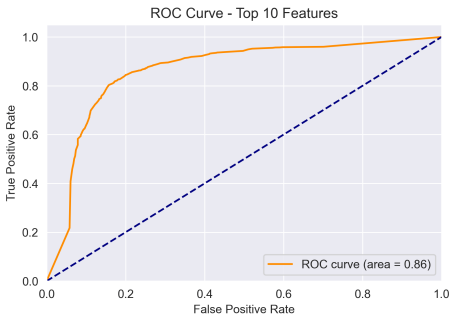
(c) Location Features - Area = 0.86



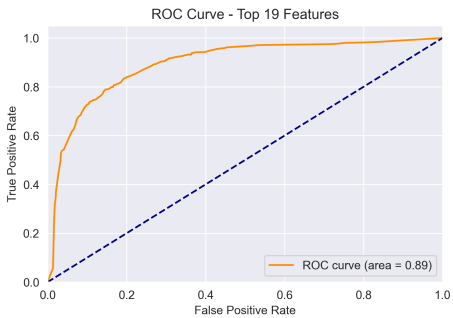
(d) RQ1 Features - Area = 0.87



(e) RQ1 Advanced Features - Area = 0.87



(f) Top 10 Features - Area = 0.86



(g) Top 19 Features - Area = 0.89

Figure 19: Seen in these figures are the ROC curves and associated areas under the curve (AUC, see also Table 3) for the classification models using the features sets as laid out in Table 1. This metric is widely used for evaluation of binary classification tasks. It depicts the True Positive Rates and the False Positive Rates of the various models, and compares the performance of the model to a 'simple' classifier that randomly assigns classes (straight blue dotted line).