UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di laurea Magistrale in

SCIENZE STATISTICHE ED ECONOMICHE

BASKETBALL ANALYTICS: THE USE OF DATA SCIENCE TO DESCRIBE AND PREDICT THE PERFORMANCE OF AN NBA TEAM

Relatore: Prof. Matteo Maria Pelagatti

Correlatore: Prof.ssa Paola Zuccolotto

Tesi di Laurea di:

Alberto Ferrario

Matr. N. 803335

Anno Accademico 2020/2021

A chi c'è sempre stato e sempre ci sarà…

Abstract

Basketball is one of the most popular sports in the world. The National Basketball Association (NBA) is the first professional league according to number of fans and players' revenues.

Data science offers reliable tools for the evaluation of the NBA teams' performances, from this comes Basketball Analytics.

This research has two main aims: on one hand it wants to give an overview of the most known and powerful metrics used to describe the performance of an NBA team, on the other it attempts to predict the outcome of regular-season games by using as predictors the average team's performances along with the data related to their Conference standings before the start of the games.

The data of six recent NBA regular seasons were used in order to carry out the analysis. The results highlighted the disparities among Eastern and Western Conference, with the teams belonging to the latter, standing out among the others according to their average performances. As regards to predictive analysis, the final selected model was an ensemble of three separately trained and tuned base learners: *Penalized Discriminant Analysis*, *Boosted Logistic Regression* and *C5.0*. The model achieved an *accuracy* of 0.6939, correctly predicting 757 game's outcomes out of 1091 available on the test set. *Elo Ratings*, although only recently applied to basketball data, turned out to be one of the most powerful tools for both describing and predicting teams' performances.

Contents

1 – Basketball and Statistics

This introductory section is structured of three parts. First, a brief history of basketball and its development across the years is given, especially focusing on the National Basketball Association (NBA), the first basketball league of North America. Then, the focus shifts on some more technical aspects of this sport, giving an overview of its evolution throughout the years, by exploring the main rules of the game and the organizational structure of the NBA. Finally, an overview of Basketball Analytics is given, explaining how it supports the basketball experts in taking the right decisions.

1.1 - From the Origins to the Modern NBA

It all began in December 1891[1], when the then thirty-year-old James Naismith, professor of physical education at the YMCA International Training School in Springfield, Massachusetts, was asked to create a sport to be practiced indoors to keep pupils in shape during the cold winter months. He who made this request, the schools Superintendent of Physical Education, Dr. Luther. H. Gulick, added that this new game should be "fair for all players and not too rough"[2].

Thus, inspired by an old game of his childhood in which you had to hit a target by throwing a stone ("Duck on a Rock") and thanks to the help of his wife, Maude Evelyn Sherman, after a few days Naismith wrote the thirteen basic rules of the game[3]. On 21st December 1891, he organized the first experimental game using a soccer ball and two baskets of peaches hanging at 10ft high: two teams of nine players each, arranged in a rectangular court, had the aim of throwing the ball into the basket of the opposing team.

[1] https://www.basketballforcoaches.com/basketball-history/
[2] https://www.biography.com/scholar/james-a-naismith
[3] https://www.usab.com/history/dr-james-naismiths-original-13-rules-of-basketball.aspx

Naismith could not be aware of this, but that day he created one of the most famous and practiced sports in the world, basketball.

https://www.nationalgeographic.com/history/article/basketball-only-major-sport-invented-united-states-how-it-was-created

After the first public game, recorded by the Springfield Republican on 12th of March 1892[4], played between teachers and students of the Naismith's school and ended 5 to 1 in favour of the latter, the popularity of the sport grew rapidly in colleges throughout the United States through the network of YMCAs (Young Men's Christian Association) and soon became of national interest. In 1898[5], the first professional league, the National Basketball League (NBL), was founded, which consisted of six teams, but was disbanded in 1904. In the same year, a basketball exhibition game was played at the St. Louis Olympics[6], but only at the 1936 Berlin Olympics was it introduced as an Olympic discipline, becoming popular all over the world.

In 1937, with the help of major sponsors of the time such as Goodyear and General Electric, the NBL was reintroduced, this time consisting of thirteen teams. Twelve years later, in 1949[7], from the merger of the NBL and the Basketball Association of America

[4] https://www.rarenewspapers.com/view/206238
[5] https://www.nationalgeographic.com/history/article/basketball-only-major-sport-invented-united-states-how-it-was-created
[6] https://www.olympedia.org/editions/3/sports/BAS
[7] https://nbahoopsonline.com/History/

(BAA), another rival professional league, the well-known National Basketball Association (NBA) was born, initially composed of seventeen teams. Since then, the number of participating franchises has undergone reductions and expansions, up to the current thirty teams (twenty-nine American and one Canadian).

In the first decades of life, the success of the NBA struggled to take off: due to financial problems, numerous franchises were forced to retire and the number of participants in the tournament was just eight teams in the 50s. In the 60s and 70s the number of teams increased again, and the league gained greater notoriety thanks also to the rivalry with the American Basketball Association (ABA), a professional league in operation from 1967 to 1976, when it was merged with the NBA.

It was only in the 80s that the NBA saw its success grow exponentially under the guidance of David Stern[8], NBA commissioner from 1984 to 2014. He was able to transform the league into a global international entertainment company, highlighting, thanks also to aggressive marketing campaigns, sports celebrities such as Larry Bird, Magic Johnson and Michael Jordan. The latter, six times NBA Finals MVP[9], as many times winner with his team of the NBA championship and icon of the Chicago Bulls, in the 90s was probably the sportsman who had the greatest impact on public opinion and who made basketball and the NBA known all over the world, also conveyed thanks to his clothing brand, “Air Jordan" by Nike.

Over the years, Stern helped to grow the NBA movement with numerous other ploys[10], including:

- the dispute of some preseason games outside the USA borders;

[8] https://www.sonicsrising.com/2014/1/30/5349474/the-legacy-of-nba-commissioner-david-stern-global-dominance
[9] https://www.nba.com/history/awards/finals-mvp
[10] https://www.nba.com/pelicans/news/ten-ways-david-stern-helped-grow-game-basketball

- the creation of the NBA All-Star weekend, which added to the usual All-Star game played between the best league's players, rookie game, skills challenge, three-point shootout and dunk contest;
- the introduction of a salary cap, to allow all teams to compete for the championship regardless of the economic availability and size of the host city;
- the creation of the "Dream Team", with which the USA participated and won at the 1992 Barcelona Olympics, fielding for the first time the strongest players in the NBA and not the youngsters of the university championship, as had happened in previous editions of the Olympic Games.

Stern also contributed significantly to the spread of the women's professional movement, as he oversaw the creation of the Women's National Basketball Association (WNBA), the NBA's female equivalent, which debuted in 1997 with its first official season. Finally, in the early 2000s, Stern collaborated in the founding and development of the NBA Development League (now the G League), the NBA's official minor league, which aims to prepare players, coaches and all the figures involved in the world of basketball, to the NBA, also acting as a research and development laboratory[11].

In the first two decades of the new millennium, the NBA has grown steadily internationally, attracting several players from all over the globe, such as the German Dirk Nowitzki, sixth all-time points scorer in the league[12], or the Italian Andrea Bargnani, who in 2006[13] became the first and to date also the only European player to be selected as the first pick in the Draft, or even the French Tony Parker, who with his three NBA championships, an NBA Finals MVP and eighteen seasons played is the second foreign player for NBA games played and the thirty-fifth overall[14]. But the European one is not

[11] https://gleague.nba.com/about/
[12] https://www.nba.com/stats/alltime/#!?SeasonType=Regular%20Season&PerMode=Totals
[13] https://www.nba.com/history/draft
[14] https://www.basketball-reference.com/leaders/g_career.html

the only continent represented in the NBA, in fact the other continents have also been and still are able to produce big stars. For Oceania, which between Australia and New Zealand has provided about thirty players, the best known is Andrew Bogut, first pick in the 2005 Draft and champion in 2015 with the Golden State Warriors. The Asian continent, despite having produced only 22 NBA players, can boast of having a player belonging to the Naismith Memorial Basketball Hall of Fame[15], Yao Ming, who with his 7 feet and 6 inches of height amazed America and contributed more than anyone else to the spread of the NBA in China. From Africa, and in particular from Nigeria, numerous players have arrived since the 90s, including authentic NBA stars, such as Hakeem Olajuwon, Dikembe Mutombo and Joel Embiid. The peak of foreign representation in the NBA roasters was reached in the 2016-17 and 2017-18 seasons, when at the beginning of the former[16] the foreign players were one hundred and thirteen from forty-four countries and territories, while the following season[17] saw the debut of one hundred and eight foreign players registered from forty-two countries and territories.

To date[18], games are broadcast in 215 countries and 49 different languages, and NBA merchandise is available in more than 125000 official retailers in 100 different countries. Since 2005, NBA teams and players have been actively engaged in social responsibility initiatives in the United States and the rest of the world, thanks to the NBA Cares project[19], which collaborates with internationally recognized youth-serving programs and is responsible for supporting the education, human development and health of children and their families in need. The NBA moreover, at the end of the 2018-19 season, with 8.76 billion dollars in revenue[20], was in third place in the ranking of

[15] https://www.nba.com/history/hall-of-fame-inductees
[16] https://pr.nba.com/nba-rosters-international-players-2016-17/
[17] https://pr.nba.com/nba-international-players-2017-18/
[18] https://careers.nba.com/our-leagues/
[19] https://cares.nba.com/
[20] https://www.statista.com/statistics/193467/total-league-revenue-of-the-nba-since-2005/

professional sports leagues by volume of revenue, behind only the National Football League (NFL) and Major League Baseball (MLB) with respectively 16 and 10.7 billion dollars in revenue[21] [22], thus ranking as one of the richest sports leagues in the world. The salaries of the players, although limited by the salary cap, which decrees the amount of money that the franchise can spend, have grown exponentially in the last thirty years, and today the NBA is the highest paid sport in the world[23]. During the 2019/20 season, a player's average annual salary was around $8.32 million, while the sum of all salaries slightly exceeded $3.65 billion[24]. Thanks to the fame gained on the courts of the NBA, stars such as Steph Curry and Lebron James sign multimillion-dollar advertising and movie contracts and are among the best-known sportsmen, as well as the highest paid in the world.

https://www.adsoftheworld.com/media/print/nba_the_finals

[21] https://www.statista.com/statistics/193457/total-league-revenue-of-the-nfl-since-2005/
[22] https://www.forbes.com/sites/maurybrown/2019/12/21/mlb-sees-record-107-billion-in-revenues-for-2019/#70ecc8935d78
[23] https://www.statista.com/statistics/675120/average-sports-salaries-by-league/
[24] https://www.statista.com/statistics/1120257/annual-salaries-nba/

1.2 - NBA Rules and Organization

From the day Naismith wrote the thirteen rules of basketball, they have been reworked and expanded over the years, up to more than a hundred current in the NBA, but the spirit and principles of the original ones are still valid today.

The most significant changes[25] concerned the number of players, the size of the court and the seconds available to conclude an action.

In particular, one of the first changes concerned the number of players on the court: initially there was no rule that set the number (as mentioned above, the first game was played in nine against nine and this because Naismith's class was composed of eighteen students), but in 1897 the number was established in five players and became the standard still adopted today. As for substitutions, however, they were not originally allowed, but over the years, the number of times a player could leave and return to the court was increased, until 1945, when the rule was established that there is no limit to the maximum number of substitutions and the number of times the same player can be changed.

The size of the court was initially not well defined and could vary from game to game: in most cases the boundaries of the court were in fact the walls of the gym where the game was played, only in 1904 the demarcation lines of the playing court were introduced, while in 1932 the halfway line was also introduced, in such a way as to eliminate stalling. The current size of an NBA court[26] is 94 feet long by 50 feet wide, while the height of the basket, which has become a cast iron rim with an open-ended nylon net since 1912, has remained 10 feet above the ground, as at the beginning.

[25] https://hooptactics.net/premium/basketballbasics/bb8rulesevolution.php
[26] https://official.nba.com/rule-no-1-court-dimensions-equipment/

The total duration of a game underwent several changes, going from the original two times of fifteen minutes to the current four times of twelve minutes each in the NBA, a duration that differs from that provided for by the FIBA (International Basketball Federation) regulations, applied in the rest of the world and which provides for four times of ten minutes each. On the other hand, with regard to the time available to conclude an action, it was not initially defined and, in this way, numerous situations were created in which the team in the lead kept possession of the ball as long as possible, to avoid giving the opponents the opportunity to recover the gap, with the consequent slowdown of the game. In 1954, the NBA established the 24-second rule to shoot, in favor of faster and more spectacular actions.

Another important change was that concerning the value of scoring: initially a goal made was awarded a point, regardless of the position from which the shot started; in 1894, free throwing was introduced for the first time, and two years later the rule was changed to give two points to a field goal made and one to a free throw made. The three-point shot, and its demarcation line, made its debut in the United States in 1961, thanks to the American Basketball League (ABL), a professional league of little following, and was then popularized in 1967 when it was adopted by the ABA. It was only in 1979 that the rule was introduced in the NBA; on the 12th of October of the same year[27], Chris Ford of the Boston Celtics scored the first three-point mark in NBA history.

Finally, this may seem strange, but for many years the figure of the coach remained only marginal: he, in fact, could communicate with the players only before the game and during the interval in the middle of the match. Beginning in 1949, however, the rule was introduced whereby the coach could interact with the players at any time during the match. Thanks to this change, the coach also acquired a potentially fundamental role

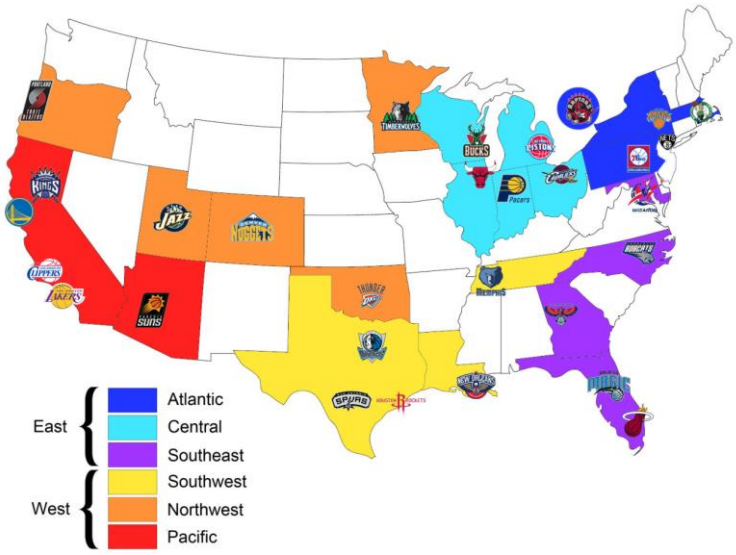
[27] https://ca.nba.com/news/nba-history-birth-evolution-3-point-line-stephen-curry-reggie-miller-ray-allen/zlqxs2380v7o1pn4oeumjhsmh

in the victory of a team and, as a result, he began to be considered by many the "sixth man" on the court. Over the years, many coaches were distinguished by their successes and the most titled were inducted into the Naismith Memorial Basketball Hall of Fame. Among them it is worth mentioning Phil Jackson[28], who with eleven NBA championships and the impressive record of 1155 wins out of 1640 games played in the regular season (0.704 of winning percentage) over twenty seasons, is considered the greatest head coach in the history of the NBA. Other legendary names are those of Red Auerbach[29], historic head coach of the Boston Celtic from 1950 to 1966, who led his team to an unrepeatable sequence of eight consecutive NBA Finals victories from 1959 to 1966, and Gregg Popovich[30], head coach of the San Antonio Spurs from 1996 to the present (with which he won five NBA championships and three awards for best coach of the year), who, with his 25-year career, is the longest-serving active head coach in the NBA and from April 13, 2019[31] also the most successful ever between regular season and playoffs, surpassing Lenny Wilks, who stopped at 1,412 victories.

As for the structure of the tournament, it has undergone numerous variations over the years. First, the current participating franchises and the division into groups date back to the 2004/05 season. The thirty teams are equally divided into two Conferences, Eastern and Western: the fifteen teams of the Eastern Conference are in turn divided into three Divisions (Atlantic, Central and Southeast), while the fifteen teams belonging to the Western Conference are divided between Southwest, Northwest and Pacific Division. The following figure shows the distribution and subdivision of the various NBA teams

[28] https://ca.nba.com/news/nba-history-birth-evolution-3-point-line-stephen-curry-reggie-miller-ray-allen/zlqxs2380v7o1pn4oeumjhsmh
[29] https://www.basketball-reference.com/coaches/jacksph01c.html
[30] https://www.basketball-reference.com/coaches/auerbre99c.html
[31] https://www.basketball-reference.com/coaches/popovgr99c.html

throughout the United States (the Toronto Raptors, the only Canadian team in the tournament, are aggregated to the Atlantic Division in the East Conference).



https://mapsontheweb.tumblr.com/image/63181001743

Unlike any other sports league outside of the United States, competing teams are always the same year after year. In fact, the NBA, as well as the NFL and MLB, does not respond directly to any national sports federation and therefore does not consider the criteria for promotion to major series or relegation to minor leagues.

An ordinary NBA season is divided into three phases: preseason, regular season, and playoffs.

The first phase begins after the summer break, in September, with the so-called training camps that allow the coach and his staff to evaluate the players (especially the rookies) and select the twelve (plus three inactive) to be included in the roster, identify the strengths and weaknesses of the team, and fine-tune the game strategy. After this phase, preseason exhibition games begin, which are often played in "non-NBA" cities or even overseas.

The last week of October is typically characterized by the first games of the regular season, which lasts until about mid-April, with a break in February to allow the All-Star Weekend to take place. Each team plays forty-one games at home and as many away games, for a total of eighty-two games. A team during the season faces four times the other four teams in its Division (sixteen games), four times six teams from the other two Divisions in its Conference (twenty-four games), three times the remaining four teams in the Conference (twelve games), and finally twice the other fifteen teams from the other Conference (thirty games). This asymmetrical structure implies that the "Strength of Schedule" (*SOS*) is variable from team to team and therefore that the calendar can be more or less demanding based on the performance of the opponents. However, since all teams face each other at least twice, the *SOS* in the NBA has a lower degree of variability than other leagues, such as the NFL or MLB[32].

Ultimately, the final phase is that of the playoffs: in the months of April and May the eight best placed teams in the Conference standings at the end of the regular season face each other to decree the two Conference champions, who then clash in June in the last games of the season, the NBA Finals, to establish the winning team of the NBA championship, which is awarded the Larry O'Brien Championship Trophy. Starting from the 2015/16 season[33], the participants in the Finals are defined exclusively by the Conference ranking: the top eight in order of win-loss records enter the next phase, while previously they were identified with a more complex system based also on the Division ranking. The playoffs proceed in both Conferences with a classic board structure: in the first round the first classified at the end of the regular season faces the eighth, the second faces the seventh and so on. The winners face each other in the second round (Conference

[32] https://www.nbastuffer.com/analytics101/strength-of-schedule-sos/
[33] https://www.nba.com/2015/news/09/08/nba-to-seed-conference-playoff-teams-by-record/index.html

Semifinals) by pairing in a classic way, as represented in the example figure below (NBA Playoffs 2016). The two winning teams face each other in the third round (NBA Conference Finals) which decrees the champions of their respective Conferences, who will face in the last round, the NBA Finals, the final act of the season.

Since the 2002/03 season[34], each round is played in the best-of-seven series, so the first team to win four games advances to the next round. In addition, starting from the 2013/14 season[35], the rounds are organized according to the 2-2-1-1-1 format: the team with the home-court advantage, that is the one best positioned in the standings between the two challengers, hosts the first, second and possibly the fifth and seventh games, the rival team instead plays the remaining games on the home court.

https://www.basketcaffe.com/nba-playoffs-2016/

The current playoff system has received several criticisms, in particular regarding the disparity between the Eastern and the Western Conference: sometimes the Eastern teams with a negative win-loss record still manage to enter the playoffs, while the Western teams with a positive record fail to enter the final stages. Nevertheless, of the nineteen

[34] http://a.espncdn.com/nba/news/2003/0208/1506023.html

[35] https://web.archive.org/web/20131030061032/http://www.nba.com/2013/news/10/23/nba-board-of-governors-format-change.ap

franchises that have won at least one NBA championship[36], ten are from the East Conference (thirty-eight total wins), nine from the West Conference (thirty-four total wins), so this disparity apparently does not seem to exist, but it will be investigated later on this paperwork.

Although all thirty current NBA franchises have participated in the Playoffs at least once, some have never managed to win the championship, such as the Phoenix Suns, runners-up of the 2021 Finals, or the Orlando Magic, while other teams have never even reached the Finals, among them the Los Angeles Clippers and the Denver Nuggets, who, respectively in forty-nine and forty-three seasons, have never managed to access the decisive series for the championship. On the contrary, the franchises with the most participations in the Finals are the Los Angeles Lakers and the Boston Celtics, respectively present thirty-two and twenty-one times and both with seventeen victories, while the Chicago Bulls, boast an absolute record: they are the only team that has participated several times in the Finals without ever losing, in six appearances they have achieved six victories, all in the 90s under the guidance of Phil Jackson and with Michael Jordan on the court.

A last important aspect that distinguishes the NBA from the other overseas leagues is the Draft: an event that takes place every year normally towards the end of June or in any case after the Finals, in which the thirty franchises have the opportunity to contract players, usually from colleges or from abroad, who meet certain eligibility criteria[37]. Since 1989[38], the Draft is organized into two sessions, in which, one at a time, all the teams make their choice. In addition, thanks to the Draft Lottery system, introduced for the first time in 1985 and modified several times over the years, teams that have not

[36] https://www.nba.com/history/season-recap-index
[37] https://www.webcitation.org/6EMRU2GS1?url=http://www.nbpa.com/sites/nbpa.org/files/ARTICLE%20X.pdf
[38] https://web.archive.org/web/20101203184544/http://www.nba.com/history/draft_evolution.html

qualified for the Playoffs and that during the season have achieved a win-loss record lower than that of the others, have a greater chance of being drawn among the first to be able to make their choice.

1.3 - Data Science and Basketball

Data science is a powerful tool to support those who, on the basis of the many information available, have to take decision. This also applies to evaluations that can be made in the context of different sports, such as basketball.

In Sports Analytics, there is a close collaboration between sport experts and Data Scientists. The former formulate the problems and questions, which allow the latter to understand the issue. Together they choose the research plan, determining which data to consider. The analyst, using his statistical skills and tools, processes the data and returns the results, which are interpreted together with experts. The final decisions are taken considering results from Data Science and knowledge and competence of experts[39].

The website *Journal of Basketball Studies*[40]by Dean Oliver, represents a milestone for Basketball Analytics. He, thanks to his studies into the importance of Pace and Possessions, the definition of Offensive and Defensive Efficiency and the development of the Four Factors of Basketball Success, is one of the most popular Basketball Analyst in the world.

Another representative Basketball Analyst is Nate Silver. In his website *FiveThirtyEight*[41], he adjusted the *Elo Rating system* in order to apply it to the evaluation of NBA teams' performances.

[39] Zuccolotto, Paola; Manisera, Marica (2020), Basketball Data Science – with Applications in R. Chapman and Hall/CRC, pp. 3-13
[40] http://www.rawbw.com/~deano/
[41] https://fivethirtyeight.com/

2 – Data Analysis

The main task of this research is to explore and analyse NBA data. This Analysis of the data was produced with the open-source edition of *RStudio*[42] (*v. 1.4.1717*), an integrated development environment for *R* (*v. 4.1.0*), a programming language and software environment[43]. It is structured into three main parts, to which correspond just as many main *R* packages.

The first section of the analysis is dedicated to the collecting and understanding of statistics and metrics related to basketball teams. To achieve this task the *dplyr* package[44] is widely used. It solved all the *data manipulation* challenges occurred prior to extract useful information from the raw datasets at disposal, resulting in few lines of code. Here the data are first carefully checked and cleaned, and then enriched with some of the most recent tools of basketball analytics, above all the *Elo Ratings*.

The second part instead, shifts the focus on *data visualization*. Here the leading package is *ggplot2*[45], a powerful tool that for creating and modifying explanatory and clear charts. The goal of this section is double: on one side, the evolution over time of the rating metric mentioned above is depicted and explored in order to find patterns in data regarding to teams' strength; on the other side, charts related to a personally developed metric, *Value of Ball Possession Differential*, are analysed in order to understand if its values are reasonable for describing the performance of an NBA team during the regular season.

Finally, the last part of the research aims to *predict* the outcome of the NBA regular season games. This goal is carried out by the *caret* package[46] that attempts to

[42] https://www.rstudio.com/products/rstudio/
[43] https://www.r-project.org/about.html
[44] https://dplyr.tidyverse.org/
[45] https://ggplot2.tidyverse.org/
[46] https://topepo.github.io/caret/index.html

streamline the process for creating predictive models. Both basics and advanced machine learning models were trained and tuned in order to improve the *accuracy* of predictions. By the development of several models, it will also be possible to identify the most important variables for the predictive task. Not surprisingly, *Elo Ratings* will turn out to be among the first predictors for importance, revealing himself to be the protagonist of the research.

The whole code along with data used for the Data Analysis is available at the following link: https://github.com/albertoferrario95/nba/upload/main. The full list of *R* packages required to perform the analysis is the following: *arm, BasketballAnalyzeR, c50, car, caret, corrplot, doParallel, dplyr, e1071, earth, forcats, ggplot2, glmnet, ipred, kernlab, klaR, MASS, mboost, mda, nnet, party, partykit, pls, plyr, pRoc, psych, ranger, rpart, runner, sringr, tidyr, xgboost*.

2.1 - Data Preparation I

Data Preparation (or Data Manipulation) is a crucial step in data analysis: basically, it is the process of acquiring, cleaning and transforming raw data prior to put them under analysis. These three sub-phases are discussed in the three following sections, respectively. First, two NBA datasets are imported and merged for the analysis purpose. Then the new dataset is carefully cleaned, replacing all the missing values and dropping useless variables. In the end, thanks to the feature engineering step, data are transformed and enriched by adding relevant metrics.

2.1.1 - Importing and Merging

The dataset used for carrying out the analysis is the NBA Enhanced Box Score and Standings (2012 - 2018) one, available on Kaggle, an online community of data scientist (https://www.kaggle.com/pablote/nba-enhanced-stats). Specifically, two files are considered: *2012-18_teamBoxScores.csv* and *2012-18_standings.csv*.

The former file contains box scores data for each of the games of the regular season played by the thirty NBA franchises from season 2012/13 to 2017/18. Each game is recorded in two rows, one for the home and the other for the away team. The total number of rows is $30 \cdot 82 \cdot 6 - 2 = 14760 - 2 = 14758$: each of the 30 teams played 82 games through 6 seasons, except for Boston Celtics and Indiana Pacers that played one less. In fact, they should have played on 16th of April 2013[47] at TD Garden Arena, in Boston, but, due to Boston Marathon bombing of the day before, the game was cancelled and not rescheduled as both teams' playoff positions were already set. That means that data relate to 7379 (7380-1) different games on six NBA seasons. The columns of the

[47] https://www.espn.com/boston/nhl/story/_/id/9175332/bruins-game-boston-postponed-blasts-mlb-stands-pat-nba-cancels-celtics-game

dataset are 123: firstly, details about date and time of the game and about officials on court are reported, then, for both home and away team, franchise and game details, traditional and advanced statistics and few ratings are recorded.

The latter file communicates standings data for each team each day in which at least a game of the regular season was played, from season 2012/13 to 2017/18. The total number of rows is 29520, that means that, as long as for each game day the complete information about all teams are reported (both they played or not that day), the number of different days of records is $29520/30 = 984$. The columns which constitute the dataset are 39, in addition to date and franchise name, a lot of details regarding rank in Conference, streak, games played until that day (both for a specific team and for the opponents all together), strength of scheduled, ratings and winning percentage are available.

After importing box scores and standings data, first thing that jumps out is the inappropriate structure of the former dataset for this kind of analysis. Basically, the information contained in rows are identical two by two, the only thing that changes is the order in which they are recorded (the former records first home team 's details and then away team's details, the latter the opposite). Therefore, is sufficient to choose the first one between each pair of rows and ignore the second one to get full details about a game. The dataset now has got half of the rows than before, 7379, and by adding season and ID variable to each game record the number of columns has grown up to 125. Finally, variables names are fixed and standardized in such a way that one starting with "H." is referred to home team, while one starting with "A." is referred to away team.

On the other hand, standings dataset has got an appropriate structure for the analysis and so does not need any adjustment. Hence you can proceed with the merging of the two datasets: first, a simple left join using date and home team name as keys is

made, then, the same operation is done but using date and away team name, finally the two new datasets are combined to get a new one.

The final dataset is made by 7379 rows and 199 columns, each row contains details about a specific game: first information about officials and game date are recorded, then, for both home and away team, basic and advanced statistics from box scores and standings information are shown.

The structure of the dataset is finally defined, now it is the time for data cleaning and for dealing with missing data.

2.1.2 - Cleaning and Missing Imputation

To begin, having a closer look to the data, you can notice that some little measures can be done.

Officials' first and last name are recorded on two different columns, hence they can be joined to a single variable combining them together, reducing the number of variables referred to them from six to three.

Minutes played are recorded both for home and away team in two different columns, but it is useless, since the total minutes played in a game are the same for both teams, and so a generic "minutes played" variable is created. Furthermore, this variable can assume only few unique values: a regular game is made by four quarters each of twelve minutes and five players are on court, that means that the total minutes played on a regular game are $4 \cdot 12 \cdot 5 = 240$. If at the ending of regular time the teams are tying, an overtime of five minutes is played. Once the overtime is finished, if teams are still in a tie you proceed with another overtime, and so on until a team overcomes the other. Any overtime adds 25 minutes ($5 \cdot 5$) to the total minutes played, hence, according to the number of overtimes, the variable in this dataset can assume only five unique values: 240

for a regular game, 265 for a 1-overtime game, 290 for a 2-overtimes game, 315 for a 3-overtimes game and 340 for a 4-overtimes game (only two out of 7379 games available ended like this). Probably due to errors in scraping data, this variable presents values that are close to the allowed values but not accurate as they should be. For this reason, the wrong values are rounded to the closest allowed value.

Before proceeding with missing imputation step, some unnecessary variables are dropped, such as team location, the hour at which the game started and the ordinal rank in championship. Also, all character variables are converted to factor, and the rank variable's levels are sorted in ascending order.

Next step is to check for missing values (*NAs*) and choose the best thing to do with them. With few lines of code, you notice that there are 1904 *NAs*, and that they all come from standings variables. Moreover, the missing values are not spread out over the rows, but they are restrained to those that refers to three specific days: 2016-01-22, 2016-11-18, 2016-11-30. In fact, going back to the original standings data, even if apparently there are no *NAs*, the standings details on these three dates are simply missing, in the sense that, during data scraping, they got lost. Due to the impossibility of providing the data requested in an automated manner, they were manually scraped from https://www.basketball-reference.com/ thanks to the "NBA Standings by Date" tool. Finally, all the new values are included in the dataset and by doing so the number of missing values is now 0.

Apparently, the missing imputation phase has come to an end, but by having a closer look, few "hidden" missing values jump out. In fact, considering the variable which refers to the third official on court (the other two are correct), there is an extra blank level that should not be there. Again, because of an error in scraping data, these values got lost. Fortunately, only two games out of 7379 have this problem, therefore first and last name

of the two missing officials are derived from the same website previously mentioned, this time using the "NBA Schedule and Results" tool, and the blank level is dropped.

At the end of the cleaning and missing imputation step, the dataset has 7379 rows and 182 columns, no missing data are present.

2.1.3 - Feature Engineering

The final step of the Data Preparation phase is feature engineering. Here this step is split into three parts: first each variable in the dataset is carefully checked and eventually fixed or recomputed, then some advanced statistics derived from literature are computed and added to the dataset, lastly two new rating variables are calculated, one originally invented as a rating-chess system but recently applied also to football and basketball[48], the *Elo Ratings*, the other personally created combining teams' and league's *Offensive Ratings*, the *Value of Ball Possession Differential*.

2.1.3.1 - Checking and Recalculating Metrics

All the 182 variables present in the dataset are checked one by one. Some have no trouble or at most their name is fixed, some, however, need to be recomputed, others are dropped. All the formulas used for the calculation of the metrics in current and following subsection can be found at https://www.nbastuffer.com/analytics-101/team-evaluation-metrics/ and https://www.basketball-reference.com/about/glossary.html, unless otherwise specified.

The following table summarises the job done. Variables highlighted in light-grey are unique values for both team and away teams, hence, they are reported just one time

[48] https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

in the dataset, on the contrary, the ones highlighted in light blue are reported two times, both for home (*H.*) and away (*A.*) team. If a formula for the metric's calculation is available, then it is recorded in the last column of the table. Here, if the formula needs only the team's statistics (and not the opponent's ones) to be computed, it is reported only once, otherwise two formulas are recorded. Also, a quick explanation is available for the most advanced metrics and for the ones essential for the analysis.

Name	**Data type**	**Description**	**Calculation**
gmID	integer	Game ID	
season	factor	Game season	
gmDate	date	Game date	
Team	factor	Team name abbreviation	(1)
Conf	factor	Team Conference	
Div	factor	Team Division	
PTS	integer	Points scored by team	
AST	integer	Assists made by team	
TO	integer	Turnovers made by team	
STL	integer	Steal made by team	
BLK	integer	Blocks made by team	
OREB	integer	Offensive rebounds made by team	
DREB	integer	Defensive rebounds made by team	

TREB	integer	Total rebounds made by team	$TREB = OREB + DREB$
PF	integer	Personal fouls made by team	
FGA	integer	Field goal attempts made by team	
FGM	integer	Field goals made by team	
FG%	numeric	Field goal percentage	$FG\% = FGM/FGA$
2PA	integer	Two points attempts made by team	
2PM	integer	Two points made by team	
2P%	numeric	Two points percentage	$2P\% = 2PM/2PA$
3PA	integer	Three points attempts made by team	
3PM	integer	Three points made by team	
3P%	numeric	Three points percentage	$3P\% = 3PM/3PA$
FTA	integer	Free throw attempts made by team	
FTM	integer	Free throws made by team	
FT%	numeric	Free throw percentage	$FT\% = FTM/FTA$

PTS1	integer	Points scored in the first quarter by team	
PTS2	integer	Points scored in the second quarter by team	
PTS3	integer	Points scored in the third quarter by team	
PTS4	integer	Points scored in the fourth quarter by team	
PTS5	integer	Points scored in the fifth quarter by team	
PTS6	integer	Points scored in the sixth quarter by team	
PTS7	integer	Points scored in the seventh quarter by team	
PTS8	integer	Points scored in the eighth quarter by team	
TS%	numeric	True shooting percentage	$TS\% = \frac{PTS}{2(FGA + 0.44 \cdot FTA)}$
FGMAST%	numeric	Percentage of field goals made by assist	$FGMAST\% = \frac{AST}{FGM}$
PPS	numeric	Points per shot	$PPS = \frac{2 \cdot 2PM + 3 \cdot 3PM}{FGA}$
TREB%	numeric	Total rebounds percentage	$H.TREB\% = \frac{H.TREB}{H.TREB + A.TREB}$ $A.TREB\% = \frac{A.TREB}{H.TREB + A.TREB}$
TOR	numeric	Turnover ratio	$TOR = \frac{100 \cdot TO}{FGA + 0.44 \cdot FTA + AST + TO}$

ASTR	numeric	Assist ratio	$ASTR = \frac{100 \cdot AST}{FGA + 0.44 \cdot FTA + AST + TO}$
AST/TO	numeric	Assist to turnover ratio	$AST/TO = \frac{AST}{TO}$
STL/TO	numeric	Steal to turnover ratio	$STL/TO = \frac{STL}{TO}$
PLAY%	numeric	Play percentage	$PLAY\% = \frac{FGM}{FGA - OREB + TO}$
FIC	numeric	Floor Impact Counter	(2)
FIC48	numeric	Floor Impact Counter per 48 minutes	$FIC48 = 240 \cdot \frac{FIC}{MIN}$
rank	ordered factor	Rank by winning percentage behind leader in Conference	
gmPlay	Integer	Number of games played by team	
gmW	Integer	Number of games won by team	
gmL	integer	Number of games lost by team	
gmBack	numeric	Number of games behind first place team in Conference	
stk	factor	Current winning or losing streak	

DayOff	integer	Number of days since last game played by team	
lastFive	Integer	Number of team wins in last five games	
lastTen	integer	Number of team wins in last ten games	
homeW	integer	Number of games won playing as home team	
homeL	integer	Number of games lost playing as home team	
awayW	integer	Number of games won playing as away team	
awayL	integer	Number of games lost playing as away team	
confW	integer	Number of games won playing against teams in same Conference	
confL	integer	Number of games lost playing against teams in same Conference	

cumPtsScore	integer	Cumulated points scored during season by team	
cumPtsAllow	integer	Cumulated points allowed during season by team	
avPtsScore	numeric	Average points scored during season by team	$avPtsScore = \frac{cumPtsScore}{gmPlay}$
avPtsAllow	numeric	Average points allowed during season by team	$avPtsAllow = \frac{cumPtsAllow}{gmPlay}$
avMOV	numeric	Average Margin of Victory during season	$avMOV = avPtsScore - avPtsAllow$
opptGmPlay	integer	Accumulated games played by opponents	
opptGmW	integer	Accumulated games won by opponents	
opptOpptGmPlay	integer	Accumulated games played by opponents of opponents	
opptOpptGmW	integer	Accumulated games won by opponents of opponents	
SOS	numeric	Strength of schedule	(3)
RPI	numeric	Relative Percentage Index	(4)

SRS	numeric	Simple Rating System	$SRS = avMOV - SOS$
pyth1391%	numeric	Pythagorean winning percentage (13.91)	(5)
wpyth1391	numeric	Pythagorean wins (13.91)	(5)
pyth165%	numeric	Pythagorean winning percentage (16.5)	(5)
wpyth165	numeric	Pythagorean wins (16.5)	(5)
off1	factor	Official 1 first and last name	
off2	factor	Official 2 first and last name	
off3	factor	Official 3 first and last name	
MIN	Integer	Total minutes played by team	$MIN = 4 \cdot 12 \cdot 5 = 240$ on a regular game, +25 for each overtime played
H.Win	factor	Target variable that says whether home team won or lost	$H.Win = \begin{cases} 1 \; if \; H.PTS - A.PTS > 0 \\ 0 \; if \; H.PTS - A.PTS < 0 \end{cases}$

(1): Codes for franchises' names are the following:

Eastern Conference		**Western Conference**	
Atlantic Division		**Northwest Division**	
BOS	Boston Celtics	DEN	Denver Nuggets
BKN	Brooklyn Nets	MIN	Minnesota Timberwolves
NY	New York Knicks	OKC	Oklahoma City Thunder
PHI	Philadelphia 76ers	POR	Portland Trail Blazers
TOR	Toronto Raptors	UTA	Utah Jazz
Central Division		**Pacific Division**	
CHI	Chicago Bulls	GS	Golden State Warriors
CLE	Cleveland Cavaliers	LAC	Los Angeles Clippers
DET	Detroit Pistons	LAL	Los Angeles Lakers
IND	Indiana Pacers	PHO	Phoenix Suns
MIL	Milwaukee Bucks	SAC	Sacramento Kings
Southeast Division		**Southwest Division**	
ATL	Atlanta Hawks	DAL	Dallas Mavericks
CHA	Charlotte Hornets	HOU	Houston Rockets
MIA	Miami Heat	MEM	Memphis Grizzlies
ORL	Orlando Magic	NO	New Orleans Pelicans
WAS	Washington Wizards	SA	San Antonio Spurs

(2): *Floor Impact Counter* (*FIC*) is a rating system developed by Chris Reina in 2007[49], similar to *Tendex* by Dave Heeren[50] and *Win Score* by David Berri[51], but it gives greater importance to assists, shot creation and offensive rebounds. *FIC* formula is the following:

$$FIC = PTS + OREB + 0.75 \cdot DREB + AST + STL + BLK - 0.75 \cdot FGA - 0.375 \cdot FTA - TO - 0.5 \cdot PF$$

[49] https://basketball.realgm.com/article/208810/The-Reina-Value
[50] https://www.nbastuffer.com/analytics101/tendex/
[51] https://www.nbastuffer.com/analytics101/win-score/

(3): *Strength of schedule* (*SOS*) represents average schedule difficulty faced by each team in the games that it is played so far or for all season[52]. Its formula considers both opponents' and opponents of opponents' winning percentage:

$$SOS = \frac{2 \cdot \frac{opptGmW}{opptGmPlay} + \frac{opptOpptGmW}{opptOpptGmPlay}}{3}$$

(4): *Relative Percentage Index* (*RPI*) measures a team's strength of schedule and how a team does against that schedule[53]. It is used to produce power ratings only considering whether a team won or lost, not taking into account the margin of victory. Its formula is the following:

$$RPI = 0.25 \cdot \frac{gmW}{gmPlay} + 0.5 \cdot \frac{opptGmW}{opptGmPlay} + 0.25 \cdot \frac{opptOpptGmW}{opptOpptGmPlay}$$

(5): *Pythagorean Winning Percentage* is a method that gives an expected winning percentage using the knowledge that team winning percentages are generally closely related to points scored and points allowed[54]. It was originally invented by Bill James in baseball and later adapted for the first time in basketball in 1993[55], by Daryl Morey. It is formulated as:

$$pyth\% = \frac{cumPtsScore^x}{cumPtsScore^x + cumPtsAllow^x}$$

Where the superscript *x* is an exponent that is empirically determined: in baseball[56] the best values were found to be around 2, instead in basketball they range from 13 to 17.

[52] https://www.nbastuffer.com/analytics101/strength-of-schedule-sos/
[53] https://www.nbastuffer.com/analytics101/relative-percentage-index-rpi/
[54] Kubatko, Justin; Oliver, Dean; Pelton, Kevin; and Rosenbaum, Dan T. (2007) "A Starting Point for Analyzing Basketball Statistics," *Journal of Quantitative Analysis in Sports*: Vol. 3: Iss. 3, Article 1, p. 17
[55] Dewan, John; Zminda, Don (1993) STATS 1993 Basketball Scoreboard (1st edition), Harperreference, p. 17
[56] https://www.mlb.com/glossary/advanced-stats/pythagorean-winning-percentage

Daryl Morey used 13.91 in his original paperwork, modern literature, on the contrary, favours bigger exponents. As both Dean Oliver[57] and John Hollinger suggest, 16.5 is a reasonable choice for the *x* value.

Pythagorean Wins are just the expected number of wins given the *Pythagorean Winning Percentage*. *Pythagorean Wins* formula is:

$$wpyth = pyth\% \cdot gmPlay$$

2.1.3.2 - Calculating New Metrics

After having inspected the available metrics in the dataset, you proceed with the computation of new ones. Again, a table helps visualizing the variables created, and the colours have the same meaning of the previous one. In addition, explanations on metrics' calculation and meaning are given, focusing on *Possessions* and *Four Factors*.

Name	Data type	Description	Calculation
TSA	numeric	True shooting attempts by team	$TSA = FGA + 0.44 \cdot FTA$
FTR	numeric	Free throw ratio	$FTR = \frac{FTA}{FGA}$
BLKR1	numeric	Block ratio 1	(1)
BLKR2	numeric	Block ratio 2	(1)
VIR	numeric	Value of Index Rating by team	(2)
(X.)POSS	numeric	Possessions	(3)
(X.)PACE	numeric	Pace	(3)
POSS	numeric	Possessions (average)	(3)
PACE	numeric	Pace (average)	(3)
FLOOR%	numeric	Floor percentage by team	(4)

[57] http://www.rawbw.com/~deano/index.html

POWER%	numeric	Power percentage by team (approximation[58] of FLOOR%)	$POWER\% = \frac{FGM + OREB}{FGA + TO}$
ORtg	numeric	Offensive Rating by team	(5)
DRtg	numeric	Defensive Rating by team	(5)
eDiff	numeric	Efficiency Differential (or Net Efficiency Rating)	$eDiff = ORtg - DRtg$
FF1O	numeric	Offensive 1st "Four Factors"	(6)
FF1D	numeric	Defensive 1st "Four Factors"	(6)
FF2O	numeric	Offensive 2nd "Four Factors"	(6)
FF2D	numeric	Defensive 2nd "Four Factors"	(6)
FF3O	numeric	Offensive 3rd "Four Factors"	(6)
FF3D	numeric	Defensive 3rd "Four Factors"	(6)
FF4O	numeric	Offensive 4th "Four Factors"	(6)
FF4D	numeric	Defensive 4th "Four Factors"	(6)
W/L%	numeric	Team's winning percentage during season	$W/L\% = \frac{gmW}{gmPlay}$
opptW/L%	numeric	Opponents' winning percentage during season	$opptW/L\% = \frac{opptGmW}{opptGmPlay}$

[58] http://www.rawbw.com/~deano/index.html

opptOpptW/L%	numeric	Opponents of opponents' winning percentage during season	$opptOpptW/L\,\% = \frac{opptOpptGmW}{opptOpptGmPlay}$

(1): *Block ratio* (*BLKR*) is a statistic personally built taking inspiration from *Block percentage* metric[59]. There are two versions of it: the former (*BLKR1*) indicates how many shots were blocked by team on the total of field goals attempted by opponents, the latter (*BLKR2*) represents the number of field goals missed by opponents due to blocks by team. Formulations are the following:

$$H.BLKR1 = \frac{H.BLK}{A.FGA} \qquad A.BLKR1 = \frac{A.BLK}{H.FGA}$$

$$H.BLKR2 = \frac{H.BLK}{A.FGA - A.FGM} \qquad A.BLKR2 = \frac{A.BLK}{H.FGA - H.FGM}$$

(2): *Value of Index Rating* (*VIR*) is an evaluation index that considers total minutes played by team. It is calculated in this way[60]:

$$H.VIR = \frac{H.PTS + \frac{3}{2}H.AST + H.STL + \frac{3}{4}H.BLK + \frac{5}{4}H.OREB + \frac{3}{4}H.DREB + \frac{1}{2}H.3PM + \frac{1}{2}A.PF - \frac{1}{2}H.PF - \frac{3}{4}(\blacksquare) - H.TO - \frac{1}{2}(H.FTA - H.FTM)}{MIN}$$

$$where\ \blacksquare = H.3PA - H.3PM + H.2PA - H.2PM$$

$$A.VIR = \frac{A.PTS + \frac{3}{2}A.AST + A.STL + \frac{3}{4}A.BLK + \frac{5}{4}A.OREB + \frac{3}{4}A.DREB + \frac{1}{2}A.3PM + \frac{1}{2}H.PF - \frac{1}{2}A.PF - \frac{3}{4}(\triangle) - A.TO - \frac{1}{2}(A.FTA - A.FTM)}{MIN}$$

$$where\ \triangle = A.3PA - A.3PM + A.2PA - A.2P$$

[59] https://www.basketball-reference.com/about/glossary.html#:~:text=BLK%25%20%2D%20Block%20Percentage%20(available,he%20was%20on%20the%20floor

[60] Imbrogno, Raffaele (2004) Statistica e Pallacanestro, p. 7

(3): *Possessions* (*POSS*) is an important metric in basketball analytics that allows to develop many others advanced statistics and to add a "per 100 possessions" dimension to the existing ones. A possession starts when a team gains control of the ball and ends when that team gives up control of it, this may happen in several ways[61]:

- making field goals or free throws that lead to the other team taking the ball out of bounds;
- missing a field goal or the last free throw and not getting the offensive rebound;
- turning the ball over.

Note that, following this definition of possession, an offensive rebound does not start a new possession, rather it starts a new play.

The evidence shows that the number of possessions in a non-overtime game is approximately the same (around 95 on average[62]) for both teams, meaning that this metric provides a useful basis for evaluating the team's efficiency. Hence, to get a more stable estimate of *Possessions*, it is calculated for both team and opponent and then the average of the two metrics is considered. The formulas are the following[63]:

$$H.POSS = H.FGA + 0.4 \cdot H.FTA - 1.07 \cdot \frac{H.OREB}{H.OREB + A.DREB} \cdot (H.FGA - H.FGM) + H.TO$$

$$A.POSS = A.FGA + 0.4 \cdot A.FTA - 1.07 \cdot \frac{A.OREB}{A.OREB + H.DREB} \cdot (A.FGA - A.FGM) + A.TO$$

$$POSS = \frac{H.POSS + A.POSS}{2}$$

[61] Kubatko, Justin; Oliver, Dean; Pelton, Kevin; and Rosenbaum, Dan T. (2007) "A Starting Point for Analyzing Basketball Statistics," Journal of Quantitative Analysis in Sports: Vol. 3: Iss. 3, Article 1, p. 2

[62] https://www.basketball-reference.com/leagues/NBA_stats_per_game.html

[63] https://www.nbastuffer.com/analytics101/possession/

Note that subsequent metrics involving *Possessions* will be calculated using the averaged formula, unless otherwise stated.

Pace (*PACE*) is a metric strictly related to *Possessions* that measures the total number of possessions a team uses in a game: in a non-overtime game the two values will be identical, whereas if at least an overtime is played, then the two values will differ. The quick formulas are[64]:

$$H.PACE = \frac{240}{MIN} \cdot H.POSS$$

$$A.PACE = \frac{240}{MIN} \cdot A.POSS$$

$$PACE = \frac{240}{MIN} \cdot POSS$$

(4): According to Dean Oliver[65], *Floor percentage* is the percentage of a team's possessions on which at least one point is scored, i.e., scoring possessions divided by total possessions. Its value can be found with the following formula:

$$FLOOR\% = \frac{FGM + 0.4 \cdot FTA \cdot [FT\%^2 + 2 \cdot FT\% \cdot (1 - FT\%)]}{POSS}$$

(5): *Offensive Rating* (*ORtg*) and *Defensive Rating* (*DRtg*) are two rating variables that measures points scored and allowed per 100 possessions, respectively. The main advantage of these two ratings is that, since the number of total possessions is approximately equal for both team in a game, they better isolate the quality of a team's

[64] https://www.nbastuffer.com/analytics101/pace/
[65] http://www.rawbw.com/~deano/index.html

offense and defence[66]. Note that in a single game a team's *ORtg* is equal to the opponent's *DRtg*. Formulas are the following:

$$H.ORtg = \frac{H.PTS}{POSS} \cdot 100 = A.DRtg$$

$$H.DRtg = \frac{A.PTS}{POSS} \cdot 100 = A.ORtg$$

(6): *Four Factors of Basketball Success* (or simply *Four Factors* or *FF*) are the answer of Dean Oliver to the question "how do basketball teams win games?"[67]. If *ORtg* and *DRtg* supply summaries of the team's overall performance on a per-possession basis, then *Four Factors* provide a breakdown of these two ratings[68]. They can be applied to both a team's offense and defence, actually resulting in eight factors (*FF1O*, *FF1D*, *FF2O*, *FF2D*, *FF3O*, *FF3D*, *FF4O*, *FF4D*). The *Four Factors* are the following:

1) *Effective field goal percentage* (*eFG%*)

$$H.FF1O = \frac{H.FGM + 0.5 \cdot H.3PM}{H.FGA} = A.FF1D$$

$$H.FF1D = \frac{A.FGM + 0.5 \cdot A.3PM}{A.FGA} = A.FF1O$$

2) *Turnover per possession* (*TO/POSS*)

$$H.FF2O = \frac{H.TO}{POSS} = A.FF2D$$

$$H.FF2D = \frac{A.TO}{POSS} = A.FF2O$$

[66] Kubatko, Justin; Oliver, Dean; Pelton, Kevin; and Rosenbaum, Dan T. (2007) "A Starting Point for Analyzing Basketball Statistics," Journal of Quantitative Analysis in Sports: Vol. 3: Iss. 3, Article 1, pp. 6-7

[67] http://www.rawbw.com/~deano/articles/20040601_roboscout.htm

[68] Kubatko, Justin; Oliver, Dean; Pelton, Kevin; and Rosenbaum, Dan T. (2007) "A Starting Point for Analyzing Basketball Statistics," Journal of Quantitative Analysis in Sports: Vol. 3: Iss. 3, Article 1, pp. 12-13

3) *Rebounding percentage* (*REB%*)

$$H.FF3O = \frac{H.OREB}{H.OREB+A.DREB} \qquad A.FF3O = \frac{A.OREB}{A.OREB+H.DREB}$$

$$H.FF3D = \frac{H.DREB}{H.DREB+A.OREB} \qquad A.FF3D = \frac{A.DREB}{A.DREB+H.OREB}$$

4) *Free throw rate* [1] (*FT Rate*)

$$H.FF4O = \frac{H.FTM}{H.FGA} = A.FF4D$$

$$H.FF4D = \frac{A.FTM}{A.FGA} = A.FF4O$$

[1]: a suitable alternative to measure this factor is to use the ratio $\frac{FTA}{FGA}$, but the team's ability to make the free throws will not be considered anymore.

Offensively, a team's aim is to minimize *Turnover per possessions* and maximize the others. Defensively, the task is to minimize *Effective field goal percentage* and *Free throw rate* and maximize the others.

The *Four Factors* can be respectively translated to four main principles needed to win a game: scoring efficiently, protecting the ball, grabbing as many rebounds as possible and getting to the foul line scoring the free throws. While they are all essential factors for winning a game, they do not carry equal weights. According to Dean Oliver[69], the approximated weights are 40%, 25%, 20%, 15%, respectively. This entails that, as one would expect, shooting well is the most desirable skill to win a game.

[69] https://www.basketball-reference.com/about/factors.html

2.1.3.3 - Elo Ratings and VBPdifferential

The ending part of this section is about two new rating metrics for basketball analytics: *Elo Ratings* and *Value of Ball Possession Differential*. Both will be included in the dataset two times, one for home and one for away team.

Elo Ratings

Elo Rating system was originally ideated by physicist Arpad Elo[70] to rate chess player, but in the last few years it became a powerful tool to rate and analyse a team's performance in many sports, from football to basketball. Nate Silver and Reuben Fischer-Baum, in their 2015[71] article on FiveThirtyEight website, explain in detail the appropriate arrangements required for calculating *Elo Ratings* for NBA teams. Also, on the same website, is possible to find the challenging project named "The Complete History of The NBA", which aim is to track each NBA franchise's performance through every game of its history using *Elo Ratings* and that continues to update with the most recent NBA games (https://projects.fivethirtyeight.com/complete-history-of-the-nba/#redskins).

Essential features of *Elo Rating* are the following:

- It is established on a game-by-game basis: after a game, the ratings updates and teams will start next game with the new updated ratings. Team's performance will be able to vary over the course of the season, as well as from a season to another.
- It depends only on the final score of the game, the home and away team's ratings before the game, and where it was played. Having home court represents a significant advantage for NBA teams, in fact the home teams win

[70] Elo, Arpad (1986) The Rating of Chessplayers, Past and Present (2nd edition), Arco
[71] https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

58-60% of their regular season games, compared to only 50% for the other American major sports (clearly before COVID-19 era)[72].

- It is part of a zero-sum system: a team will gain *Elo* points after winning a game, the opponents will lose the same number of points by losing. An upset win or a wide margin victory will mean a larger points' gaining.

The way it operates is straightforward: at the beginning[2], all teams start at a rating of 1500 Elo points. After a game is played, the winning team gains Elo points updating its rating, while the defeated one lose the same amount of points. The more the game result is unexpected or the margin of victory is ample, the larger this quantity is. Moreover, *Elo Rating* does not reset at the beginning of the season, conversely it carries over a portion of itself from one season to the next, since good teams usually tend to stay good or at most gradually decline.

[2]: It is necessary to choose a starting point from where to calculate Elo Ratings, here the first game of 2012/13 season is selected for this purpose.

Specifically, the exact updating formula for *Elo Ratings* is the following[73]:

$$Elo_{i+1} = Elo_i + k \cdot (S - E)$$

$$with:\ Elo_i = team's\ Elo\ rating\ before\ playing\ game$$

$$Elo_{i+1} = team's\ Elo\ rating\ after\ playing\ game$$

$$k = 20\,\frac{(MOV^w + 3)^{0.8}}{7.5 + 0.006 \cdot EloDiff^w}$$

$$S = \begin{cases} 1\ if\ team\ wins \\ 0\ if\ team\ loses \end{cases} \qquad E = \frac{1}{1+10^{-\frac{Elo_i - Elo_i^O + 69}{400}}}$$

$$MOV^w = winning\ team\ PTS - losing\ team\ PTS$$

[72] Chang, Wesley; Ran, Michael; Smith, Gary (2021) The Impacts of Home-Court Advantage in the NBA, p. 2

[73] https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

$$EloDiff^{w} = winning\ team\ Elo_i - losing\ team\ Elo_i$$

$$Elo_i^o = opponent's\ Elo\ rating\ before\ playing\ game$$

k is a moving constant, dependent on both difference of ratings before the game and point spread, that controls how rapidly the ratings react to the result of the game. S is a state variable indicating the game winner. E is the team's expected win probability which depends on ratings of the two contestants and incorporates home-court advantage, in fact 69 *Elo* points are added to home team's rating before the game.

Finally, about year-to-year carry-over, assuming that after last game of the season a team's $Elo = R$, at the beginning of next season its *Elo Rating* will be:

$$(0.75 \cdot R) + (0.25 \cdot 1505)$$

Historically, NBA teams' average *Elo rating* is 1500, although it can vary moderately from year to year based on league's average performance. More than 90%[74] of teams are placed between 1300 (fairly awful) and 1700 (first-class) *Elo* points, but it can happen that tremendous or deplorable teams may fall outside the range.

Value of Ball Possession Differential

Value of Ball Possession Differential (*VBPdiff*) is a new metric that compares a team's *Offensive Rating* (*ORtg*) and league's average *Offensive Rating*. The former, previously introduced, measures points scored per 100 possessions, the latter, also known as *Value of Ball Possession* or *VBP*[75] (hence the metric's name), measures the league's average points scored per 100 possessions (i.e. the average *ORtg*) on a cumulative day-by-day basis, calculated for each season.

The formula of *Value of Ball Possession Differential* is easily understandable:

[74] https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/
[75] https://www.nbastuffer.com/analytics101/value-of-ball-possession/

$$VBPdiff = ORtg - VBP$$

Thus, a positive *VBPdiff* value implies that, in a specific game, the team performed better than league's average until that game, according to points scored per 100 possessions. On the contrary, a negative value means that it performed worse. Furthermore, by taking the difference of the two quantities, the interpretation of the magnitude of *VBPdiff* is straightforward: the bigger it is the more the team moves away from the average. For example, if after a game a team's *VBPdiff* equals to -2.5, this entails that it scored 2.5 points per 100 possessions less than what league scored on average until that game.

The key idea in the building of this metric is to rate a team's performance comparing it to the average league's performance. By looking at the *VBPdiff* values through the season, you can see how the team under analysis is performing over a given period of time.

Finally, to better understand how to calculate *VBPdiff*, a quick example is reported below. Note that the following are not real records, they were created just for explanatory purpose.

gmDate	H.Team	A.Team	H.ORtg	A.ORtg	VBP	H.VBPdiff	A.VBPdiff
2015-10-27	CLE	ATL	103.42	102.78	102.66	0.76	0.12
2015-10-27	BOS	NY	105.1	99.34	102.66	2.44	-3.32
2015-10-28	LAC	MIL	117.96	105.84	106.176	11.784	-0.336
2015-10-28	HOU	LAL	101.9	96.66	106.176	-4.276	-9.516
2015-10-28	DEN	OKC	108.72	120.04	106.176	2.544	13.864

2015-10-27:

$$VBP = \frac{103.42+102.78+105.1+99.34}{4} = 102.66$$

2015-10-28:

$$VBP = \frac{103.42+102.78+105.1+99.34+117.96+105.84+101.9+96.66+108.72+120.04}{4+6} = 106.176$$

$$H.VBPdiff = H.ORtg - VBP \qquad A.VBPdiff = A.ORtg - VBP$$

More details and some explanatory plots about *VBPdiff* are available in the Descriptive Analysis section below.

At the conclusion of the Data Preparation phase, the dataset has 7379 rows and 208 columns. Checking one last time for missing values, it turns out that there is still 835 of it. This may sound surprising, but having a closer look, one can see that they are not *NAs*, while instead they are *NaNs* (*Not a Number*). This is due to the absence of data for computing some metrics and relates only to the first game of the season for each team. For example, the *opptW/L%* variable, necessary to build *SOS*, *RPI* and *SRS* metrics, is defined as the ratio between opponent's game won and game played, but at the very first game of the season, no one but the two teams on court have played or even won. This implies that the ratio will assume a value of $\frac{0}{0}$, hence the *NaN*. That is no big deal, since the rows that relates to games where at least a team takes the court for the first time in the season will be removed from the dataset (and so also the *NaNs*) during the Forecasting phase.

It is also noted that the variables belonging to the actual dataset can be used for a Descriptive Analysis as done in the section below, but almost all of them are totally useless for the predictive task of this paperwork. Therefore, in the Forecasting section, another Data Preparation phase will be necessary to extract useful information.

2.2 - Descriptive Analysis

Descriptive Analysis is a statistical tool that permits to summarise and discover patterns in data by using tables and charts. The aim of the following section is to visualize and analyse data related to the two metrics introduced in the above section for last: *Elo Rating* and *VBPdiff*. As regard the first, several plots about its evolution over time are depicted, comparing different Divisions and Conferences. About the second instead, firstly a wide look at the league's averages is given, then, choosing randomly a season, the focus shifts to exploring patterns between the metric values and the teams' ranks at the end of the season.

2.2.1 - Elo Ratings

Elo Rating, introduced in the previous section, is an innovative zero-sum rating system used to track NBA teams' performance across seasons. This section has multiple purposes: firstly, wants to show the evolution of teams' *Elo Ratings* over the period of time under analysis; next tries to discover if there is any evidence of disparity between different Conferences and Divisions according to the rating values; lastly a basic chart is used to unveil the efficacy of *Elo Ratings* for predicting the outcome of a game and to compare them with simple win-loss records.

All the *Elo Ratings* shown in the charts below has to be intended as the metric's values before playing the game. In fact, the differential between home and away teams' rating is the only information available (as well as who has home court) before the game starts since the margin of victory is clearly defined at game ending.

Figure 1 displays day-by-day Eastern Conference teams' *Elo Ratings* from 2012/13 to 2017/18 season. The fifteen teams are split into three panels according to their

Division. The solid black line represents the average rating value between teams of the same Division for each season.

Figure 1.a refers to Atlantic Division: after a broadly balanced first season, differential in *Elo Ratings* jumps out immediately. Toronto Raptors and Boston Celtics, starting from 2013/14 season and second half of 2014/15 season, respectively, are constantly placed above the Division averages. Philadelphia 76ers, on the other hand, are almost always below average, even if, during the last regular season under analysis (in which they ranked 3rd in Conference), their *Elo Rating* experiences a sudden growth, placing them above seasonal Division average.

Central Division scenario instead, is not as clear as in the previous Division: as shown in figure 1.b, there is no team that seems to outperform others over the six examined seasons. Indiana Pacers place above all the other teams in the first two seasons, but they quickly regress to Division average, and they do not manage to move from there. On the contrary, Cleveland Cavaliers, after the first two disastrous seasons, rapidly gain *Elo points* during an excellent 2014/15 season, and keep placing above others during the following three seasons that ended with three NBA Finals and a league's championship title. Milwaukee Bucks instead, except for a 2017/18 season slightly above average, collocate always below Division's average performance, with an awful 2013/14 season conducted constantly under 1400 *Elo points*.

Ultimately, figure 1.c explains the Southeast Division scenario: as happens in Central Division, there is no team persistently above the others, however the disparities between strong and weak teams are more obvious here. Miami Heats lead 2012/13 and 2013/14 seasons with an outstanding rating regularly between 1650 and 1750 points and reaching both times the NBA Finals (winning the former), before a sudden decline the following season. Conversely, Orland Magic, which never achieve to go over average

for more than few days, conduct the first three seasons with a dreadful *Elo Rating* between 1300 and 1400, ranking in Conference at the end of the regular seasons 15th, 13th and 13th, respectively.

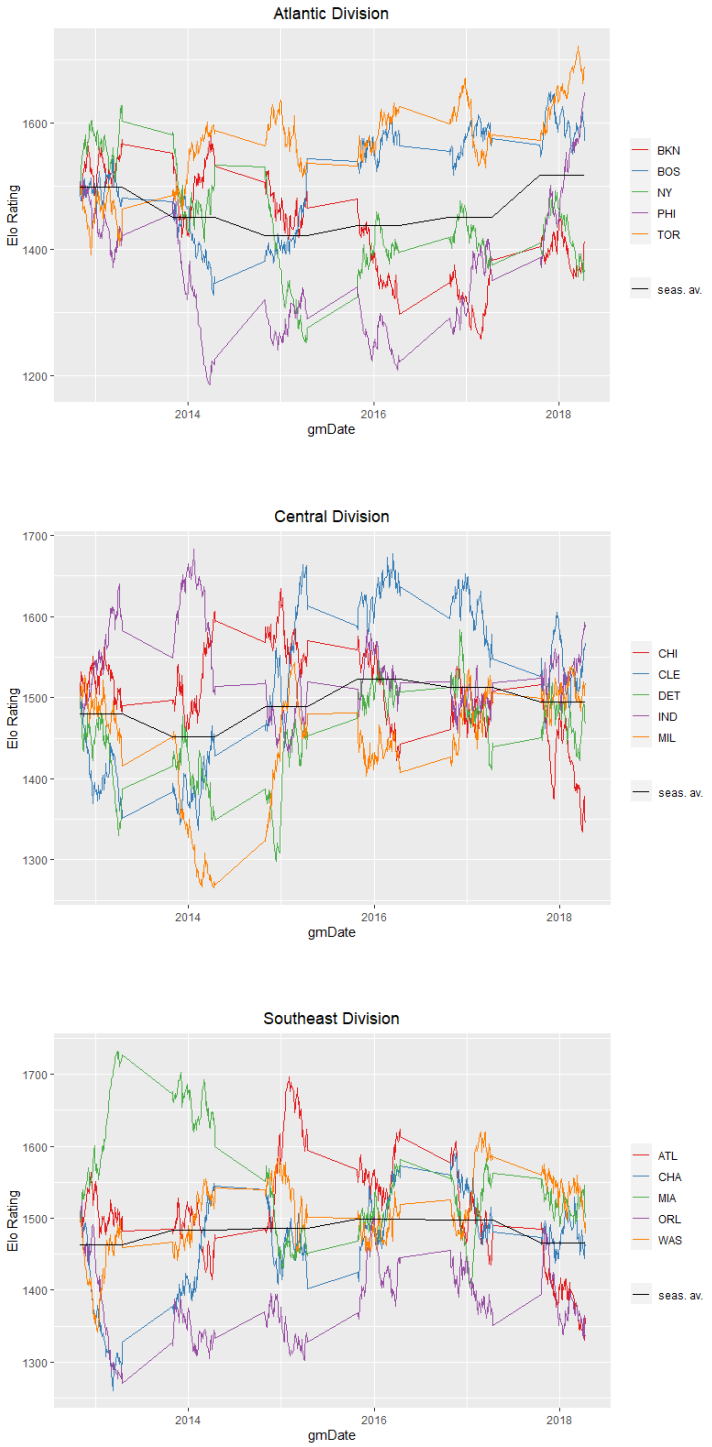


Figure 1. Eastern Conference *Elo Ratings* over seasons by Division (1.a Atlantic Division, 1.b Central Division, 1.c Southeast Division)

In the same way as for Eastern Conference, figure 2 describe the evolution of Western Conference teams over six regular NBA seasons, according to their *Elo Ratings*. Once again, teams are grouped by Division and separately analysed. The solid black lines depicted in both three charts constitute the seasonal Division averages for the *Elo Rating*, exactly as before.

Figure 2.a relates to Northwest Division: starting from the first season, two teams stand out among the others, Denver Nuggets and Oklahoma City Thunder. The former, after a firm 2012/13 season concluded at the 3rd place in Conference, during the next season, begin a merciless decline, that bring them to stay constantly under seasonal averages. The latter, after ranking 1st in Conference on the first season, keep their *Elo Rating* strictly above average and are able to lead 2012/13, 2013/14 and 2015/16 seasons above all other Division teams. Other three teams appear to hover around averages over seasons, not following any specific pattern. An interesting feature of this chart is the trend of 2017/18 season: all the teams' ratings look compacted around the Division average (which slightly grew up compared to previous seasons), meaning that, at least for that season, there were not wide disparities between teams.

Moving to Pacific Division (figure 2.b), the scenario is totally in contrast with the previously analysed cases. Here, the five teams' *Elo Ratings* seems following their paths apart, crossing only a very few times. Golden State Warriors and Los Angeles Clippers outperform the remaining teams across all the seasons, placing persistently above Division average. Precisely, Clippers dominate 2012/13 and 2013/14 season, while Warriors take the lead of the Division (and of the Conference) during the next four seasons, reaching the remarkable goal of more than 1800 *Elo Points* several times during 2015/16 season and winning three out of four NBA Finals played. Conversely, Sacramento Kings keep their *Elo Rating* constrained between 1400 and 1500, never going

over their Division average and reaching a maximum of more than 500 points less than Warriors' rating. Finally, Los Angeles Lakers and Phoenix Suns, despite a pair of decent seasons slightly above average, have an overall performance much lower than Warriors and Clippers.

In conclusion, figure 2.c shows the Southwest Division ratings' evolution. The scenario is quite similar to the one seen in the Pacific Division: San Antonio Spurs and Houston Rockets dominate other teams across almost all of the seasons, the former by almost reaching 1800 *Elo points* and leading two consecutive seasons between 1700 and 1800 points, the latter by having an outstanding second half of 2017/18 season, which bring them to an impressive step forward compared to the other Division teams. However,2014/15 and 2015/16 seasons seem quite balanced, aside from two clear exceptions: in the first season, while other teams hover around a good 1600 *Elo Rating* value, New Orleans Pelicans place constantly below them; in the second one, all the teams shift to roughly 1500 points, except for the Spurs, that never go under 1650 points.

Figure 2. Western Conference *Elo Ratings* over seasons by Division (2a. Northwest Division, 2b. Pacific Division, 2c. Southwest Division)

After having explored how the *Elo Ratings* evolve within each Division, it seems clear that all teams are not created equal. Also, disparities between Eastern and Western Conference and between Divisions are evident. The charts below have the goal to highlight these differences for the seasons under analysis.

To begin, *Elo Ratings* densities are displayed (figure 3), first on a per-season basis, then grouped by Division. A normal shape of densities would suggest that there are no big disparities between teams and that they are kindly well-matched, but here, no distribution looks normal-shaped.

The first chart (figure 3.a) does not differentiate for Conference, just focuses on the ratings' distributions across seasons: if the first season approaches vaguely to a normal distribution with a single peak, then the second and the third one feature two peaks corresponding to values slightly below and above 1500 points. Last three seasons instead, have a more waved shape with several peaks, showing that few teams clearly outperform the others (peaks corresponding to 1700-1800 points) and that some place below average (peaks corresponding to 1400 points).

The second chart (figure 3.b) does not take accounts of the seasons anymore, while instead it shows densities of the Divisions. The first three Divisions (blue tones) compose the Eastern Conference, the remaining ones (red tones) constitutes the Western Conference. It is clear that eastern teams tend to hang around the average rating (i.e. 1500), while many western teams place above average. Furthermore, having a closer look within Conferences, Atlantic Division produce lower ratings than Central and Southeast Division, on the other side, Pacific and Southwest Division show high densities corresponding to high ratings, while Northwest Division is quite balanced and does not display any high rating. Also, the Pacific Division's scenario looks interesting: despite of producing much more ratings between 1700 and 1800 than the other Divisions, and hence

the so-called “super teams”, it has a peak corresponding to 1400 points, meaning that there are large disparities between teams belonging to this Division.

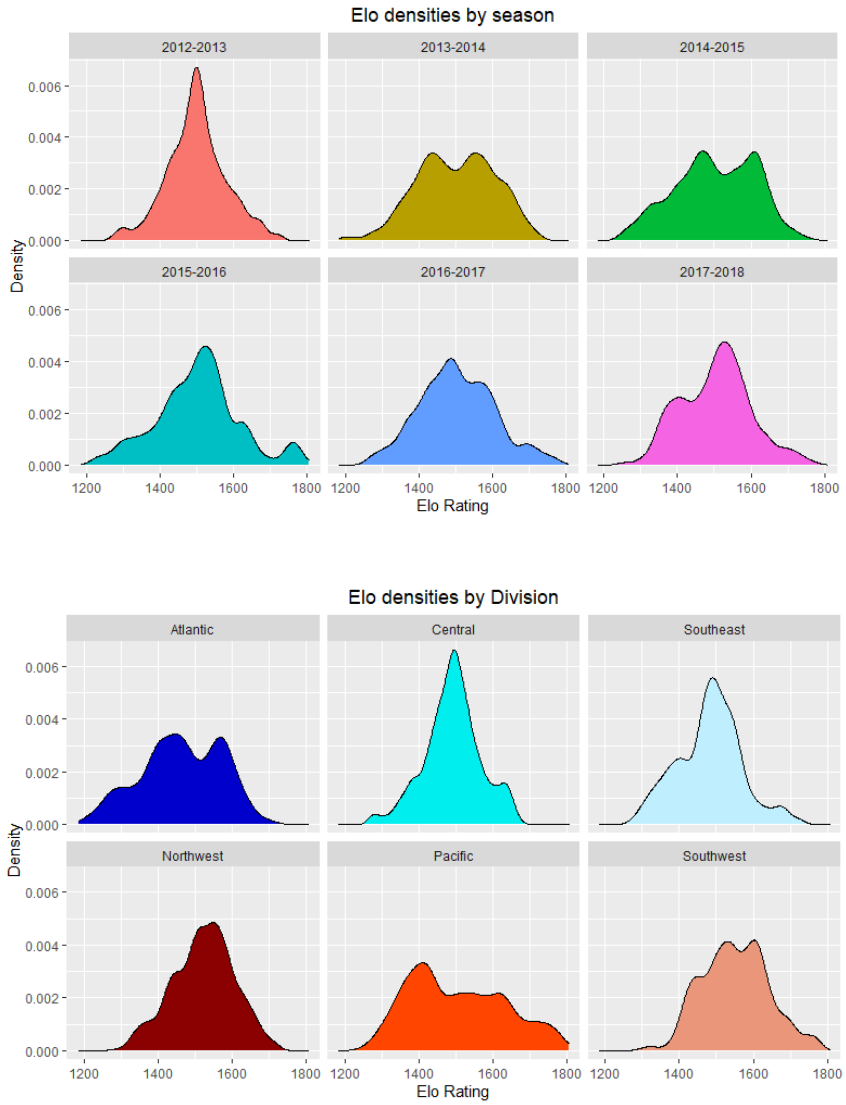


Figure 3. *Elo Ratings* densities by season (3.a) and by Division (3.b)

A clearer explanation of the disparity between Conferences is given in the two figures below.

The former chart (figure 4) compares best teams from each Division. Within each Division, the best team is chosen according to its average *Elo Rating* over the seasons under analysis. The selected teams are: Golden State Warriors from Pacific Division

(av. $Elo = 1667.375$), San Antonio Spurs from Southwest Division (av. $Elo = 1653.176$), Oklahoma City Thunder from Northwest Division (av. $Elo = 1605.9$), Toronto Raptors from Atlantic Division (av. $Elo = 1568.229$), Miami Heats from Southeast Division (av. $Elo = 1550.785$) and Indiana Pacers from Central Division (av. $Elo = 1533.945$). Note that, the six selected teams are not the first six league's teams based on overall average, the Pacers in fact, place only 10^{th} in this ranking and they are preceded by four western teams. By looking at the chart, it is evident that western teams (red tones) generally outperform the eastern ones (blue tones), even if the margin of differential between teams vary from season to season. The Heats place above all the teams during the first season, but after a decent second seasons, they rapidly decline to the bottom. The Pacers have an analogous behaviour, but the first two seasons are not as outstanding as for the Heats. The Raptors, on the contrary, show a different behaviour, starting at the bottom, but slowly growing across seasons, taking the lead on the very last days of the 2017/18 season. The Thunder instead, after two great seasons, place at the same level of eastern teams during the remaining seasons. Finally, the Warriors and the Spurs, despite not emerging during the first two seasons, stand out among the others across the next three seasons, however, during the vary last season, the latter experience a quick decline, the former instead, keep on a good level.

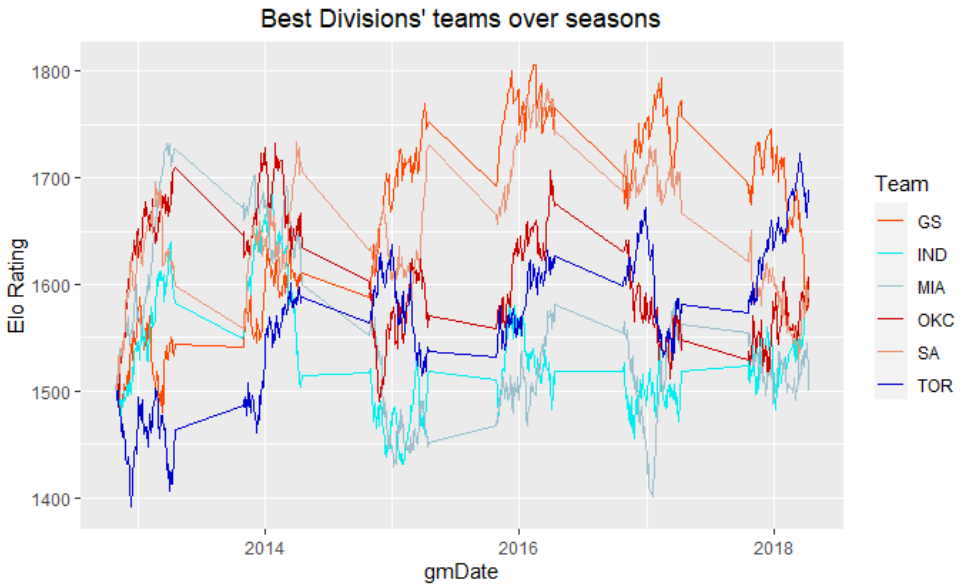


Figure 4. *Elo Ratings* over seasons of the best teams from each Division (based on their average)

The latter plot (figure 5) relates to *Elo* seasonal averages calculated for each Division separately. Once again, the disparities between Divisions and Conferences are highlighted. The western teams (dashed red tones lines) outperform the eastern ones (solid blue tones lines) across all the seasons. Southwest Division's teams have on average the highest *Elo Ratings* across the six seasons, except for the last one, when Northwest Division's teams take the lead. Pacific Division, despite having two excellent teams with an overall average greater than 1600 points (the Warriors and the Clippers), hang around 1500 points, due to the awful ratings of the other teams. Atlantic and Southeast Division seem producing the worst ratings on average, the former places just below 1425 points during 2014/15 season (against the best overall performance by Southwest teams with almost 1600 points on average), the latter does not manage to go over 1500 points neither for a season. Finally, Central Division's teams have a chequered behaviour, alternating unsatisfactory and slightly above-average seasons. Note also that,

starting from 2015/16 season, Atlantic and Northwest Division present a growing trend, while the remaining tend to slowly decline.

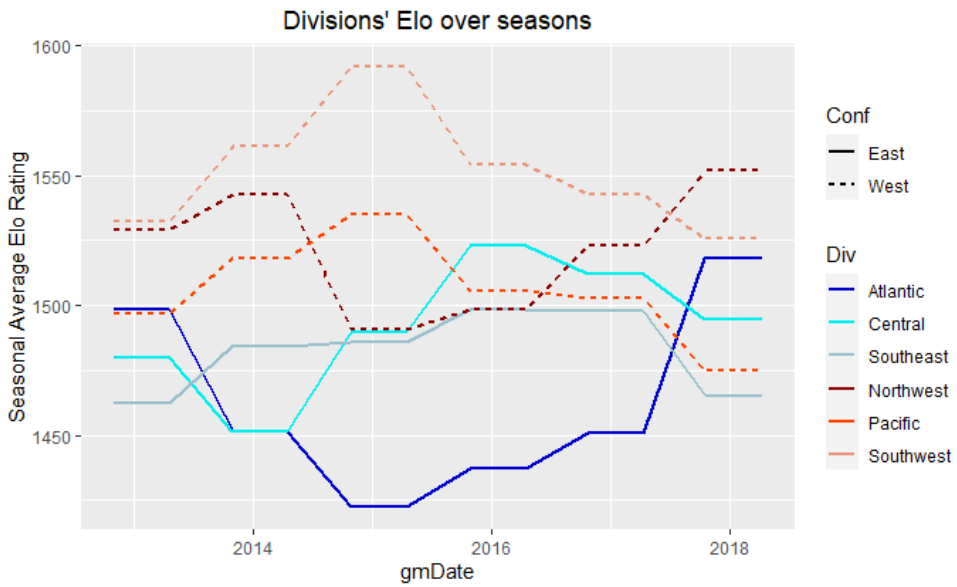


Figure 5. Seasonal averages of *Elo Ratings* for each Division

Before to move to further analysis, a little preview of forecasting analysis is showed up. A big advantage of *Elo Rating* is that, before the game starts, details on the two teams' strength are already available. A team with lower rating than opponents will start the game as underdog and by winning, it will gain more points than what it would have gained starting as favourite team. Hence the natural question that comes out is: "There is any relationship between having a greater *Elo Rating* than opponent before the game and winning it?". To answer this question, both home and away teams' *Elo Rating* before the game is calculated and the difference between the two is considered: if *Elo* differential is negative, it means that the home team is underdog, elsewhere if it is positive, the home team is the favourite. A state variable is then generated, assuming values of 1 in case the home team is the favourite, 0 otherwise. Hence, a contingency table is built between the just mentioned state variable and the outcome of the game

(assuming values of 1 in case home team wins, 0 otherwise). The results are reported in a simple pie chart represented in figure 6.a: the outcome of 64.91% of the games present in the dataset were correctly predicted, representing a considerable improvement to only using the information about home court advantage (almost 60% of winning for home teams). Finally, this method is compared to the one that takes simple win-loss record before the game for predicting its outcome. Figure 6.b shows that 64.07% of the games' results were correctly predicted. Therefore, there is an improvement in the *accuracy* of predictions by using a more sophisticated *Elo Rating system* than a simple win-loss record, even if that improvement is not large as one would expect.

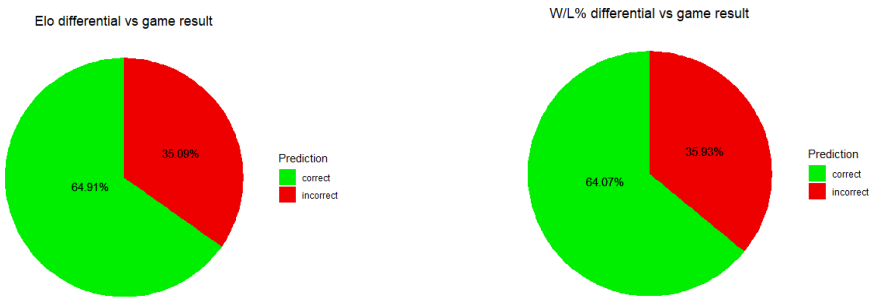


Figure 6. Predicting outcome of a game using *Elo* differential (6.a) and win-loss record (6.b)

2.2.2 - Value of Ball Possession Differential

Value of Ball Possession Differential (*VBPdiff*) is a new metric introduced in detail in the Feature Engineering section. It is the simple difference between a team's *Offensive Rating* (i.e. points scored per 100 possessions, shortened to *ORtg*) and the league's average *Offensive Rating* (also known as *Value of Ball Possession*, shortened to *VBP*). The charts below have the purpose to show how the *VBP* evolves across seasons and to highlight the relationship between *VBPdiff* and the teams' rankings in Conference at the end of the regular season.

To begin, time series of the *VBP*'s evolution from 2012/13 to 2017/18 seasons are displayed in figure 7. *VBP* is calculated on a cumulative day-by-day basis for each season, and it resets when a new regular season starts. Hence, the first values of *VBP* are always unstable and not significative, due to the very few games available to compute the league's average. As the season progresses, more games, and therefore more data, are available to calculate the metric, and so the estimates become more stable and approach to the overall average. By construction indeed, the very last value of *VBP* represents the league's overall average (depicted with a red horizontal line), since it incorporates details about all the games played over that season.

Interestingly, *VBP* seems to approach to its final value from below, growing more or less slowly but constantly over the course of the season, with the exception of 2014/15 season, when it starts with an upward trend, but after a period of stability, it falls to the average value in the second half of the season. A straightforward interpretation of this phenomenon is that the league's teams, on average, tend to increase their scoring (per possession) ability over the season. There might be many reasons why this happens, including the coaches' choice to make their teams more offensive over season, possibly to adjust the winning record, or the fact that the unity of a team increase over time, and so the players are able to improve their performance and scoring ability after a reasonable number of games.

Finally, the league's overall average appears not to follow any specific pattern from a season to another, even if it went from the value of 106-107 points scored per 100 possessions during the first four seasons, to 108-109 points per 100 possessions over the last two, resulting in a general improvement in the league's ability to score points per possession.

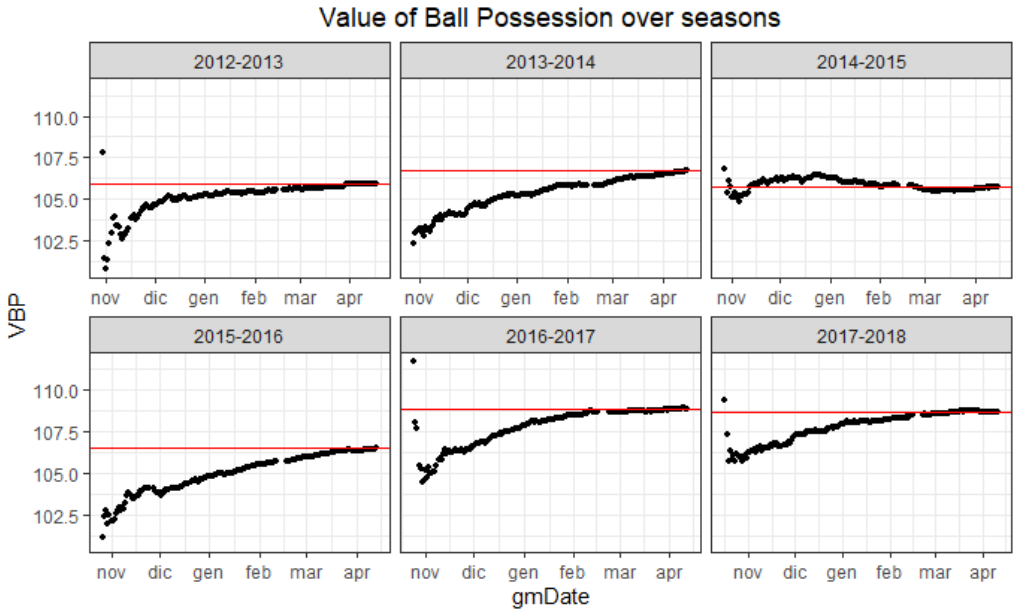


Figure 7. League's average *Offensive Rating* (*VBP*) evolution across seasons

Now, the analysis shifts the focus on *VBPdiff* that can be easily computed just by knowing the values of *ORtg* and *VBP* for each game. As previously stated, *VBPdiff*'s task is to rate a team's performance comparing it to the average league's performance. Hence, this metric not only differentiates between above-average and below-average teams, but it also measures the distance from the average performance. *VBPdiff* can be split in two parts, one for the games played as home team and one as away team, therefore they can be analysed separately.

Selecting as an example the 2015/16 season, a bubble plot related to teams' average *VBPdiff* is depicted in Figure 8 thanks to the *BasketballAnalyzeR* package[76]. The light-blue bubbles represent teams from Eastern Conference, while the red ones stand for teams from Western Conference. The number at the side of the bubble is the team's ranking in Conference at the end of the regular season. Seasonal averages of home and

[76] Marco Sandri (2020), The R package BasketballAnalyzeR, in: Zuccolotto P. and Manisera M., Basketball Data Science – with Applications in R. Chapman and Hall/CRC, Chapter 6.

away *VBPdiff* are represented on the x- and y-axes, respectively. Finally, the size of the bubble stands for the average *Elo Ratings* achieved over season rescaled between 0 and 100 (corresponding to the minimum and the maximum average *Elo* value, respectively). The two straight lines represent the league's average values of the two *VBPdiff* variables and create four quadrants. Teams placed in the top-right quadrant are the best according to *VBPdiff*, having high values both for home and away games. Conversely, the ones standing in the bottom-left quadrant are the worst, placing below average according to both variables.

By looking at the rankings of the teams located at the endpoints of the plot a clear pattern jumps out: having a high (low) average *VBPdiff* over season results in having a good (bad) final ranking. Even *Elo Ratings* seem to confirm this hypothesis, growing progressively from the bottom-left to the top-right of the chart. For both Conferences, teams placed at the top and bottom ranks at the end of the season move away from the other, locating at the endpoints of the plot. There are some exceptions, of course: in Western Conference, Oklahoma City Thunder, ranked 3rd, have greater average *VBPdiff* both for home and away games than San Antonio Spurs, ranked 2nd; besides, Houston Rockets, despite having an average value for the home *VBPdiff* and the third average away *VBPdiff* of the league, ranked only 8th at the end of the regular season. In Eastern Conference instead, Atlanta Hawks which ranked 4th, surprisingly place below home and away average *VBPdiff* values. All these contradictions raise because *VBPdiff* is clearly not a perfect rating metric: it measures the team's offensive performance, ignoring the defensive one. Further analysis could be focused on the aim to find a metric that also incorporates this aspect, producing better estimates of the teams' strength.

Two additional considerations can be done: there seems to be a positive trend in the distribution of teams, which are distributed along the diagonal going from bottom-left

to top-right, since no teams appear to perform good according to a variable and bad according to the other one, with average teams placing in the centre of the plot; western teams (except for the first and last rankings) seem having higher *VBPdiff* values on average compared to the eastern ones, highlighting one again the disparities between the two Conferences.

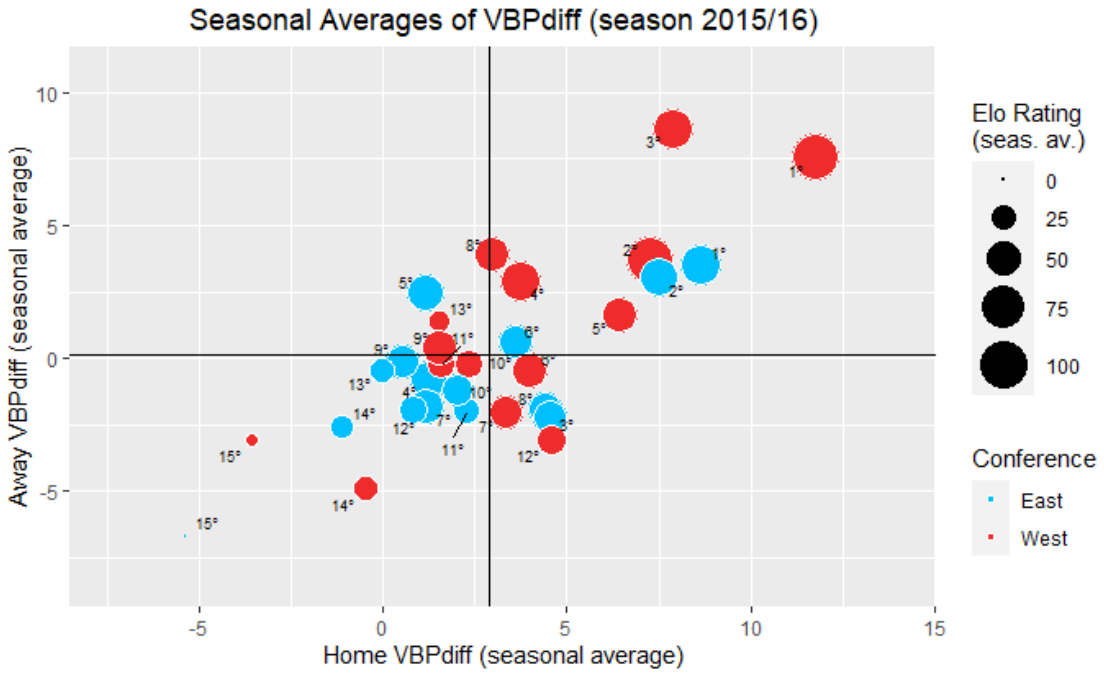


Figure 8. Home and away average VBPdiff over 2015/16 season

Ultimately, this section ends with a focus on the teams that placed 1st, 2nd, 14th and 15th in their Conferences during 2015/16 season. They are Cleveland Cavaliers, Toronto Raptors, Brooklyn Nets and Philadelphia 76ers from Eastern Conference, and Golden State Warriors, San Antonio Spurs, Phoenix Suns and Los Angeles Lakers from Western Conference, respectively. The evolution of *VBPdiff* for both home (blue bars) and away (orange bars) games is depicted in figure 9. The blue and the orange straight lines stand for home and away average *VBPdiff* over season. Figure 9.a and figure 9.b illustrate the Eastern and the Western Conference scenario, respectively. There are some

commonalities between the two charts: the two top teams ended almost all the games with a high *VBPdiff*, and the few times they achieved a negative value were often on an away game; the bottom two teams instead, collected more negative *VBPdiff* values than positive ones over the season, and the few times they achieved a positive *VBPdiff*, its absolute value was not large. In the Eastern Conference, the Cavaliers and the Raptors had almost the same average values (with the first slightly better in both), the former reached a maximum home *VBPdiff* value of 36.9 on January, the latter achieved the best away *VBPdiff* value of 31.13 at the end of February on the Memphis Grizzlies' court. On the contrary, the Nets, despite having average values slightly below zero, went under -20 *VBPdiff* value more than once over season, the same happened for the 76ers, which nevertheless had very low averages. In the Western Conference, the Warriors overcame the Spurs regarding to seasonal averages, and achieved a *VBPdiff* greater than 30 three times over the season. The Spurs instead, despite having good averages, collected few high negative values across the season, especially for the away games. Ultimately, the scenario for the last two teams is analogous to the one seen in Western Conference, with the Suns and the Lakers placing below zero both for home and away *VBPdiff*.

Hence, for both conferences, the previous hypothesis is once again confirmed: having a high (low) seasonal average *VBPdiff* both for home and away games, results in achieving a good (bad) rank in championship at the end of the season.

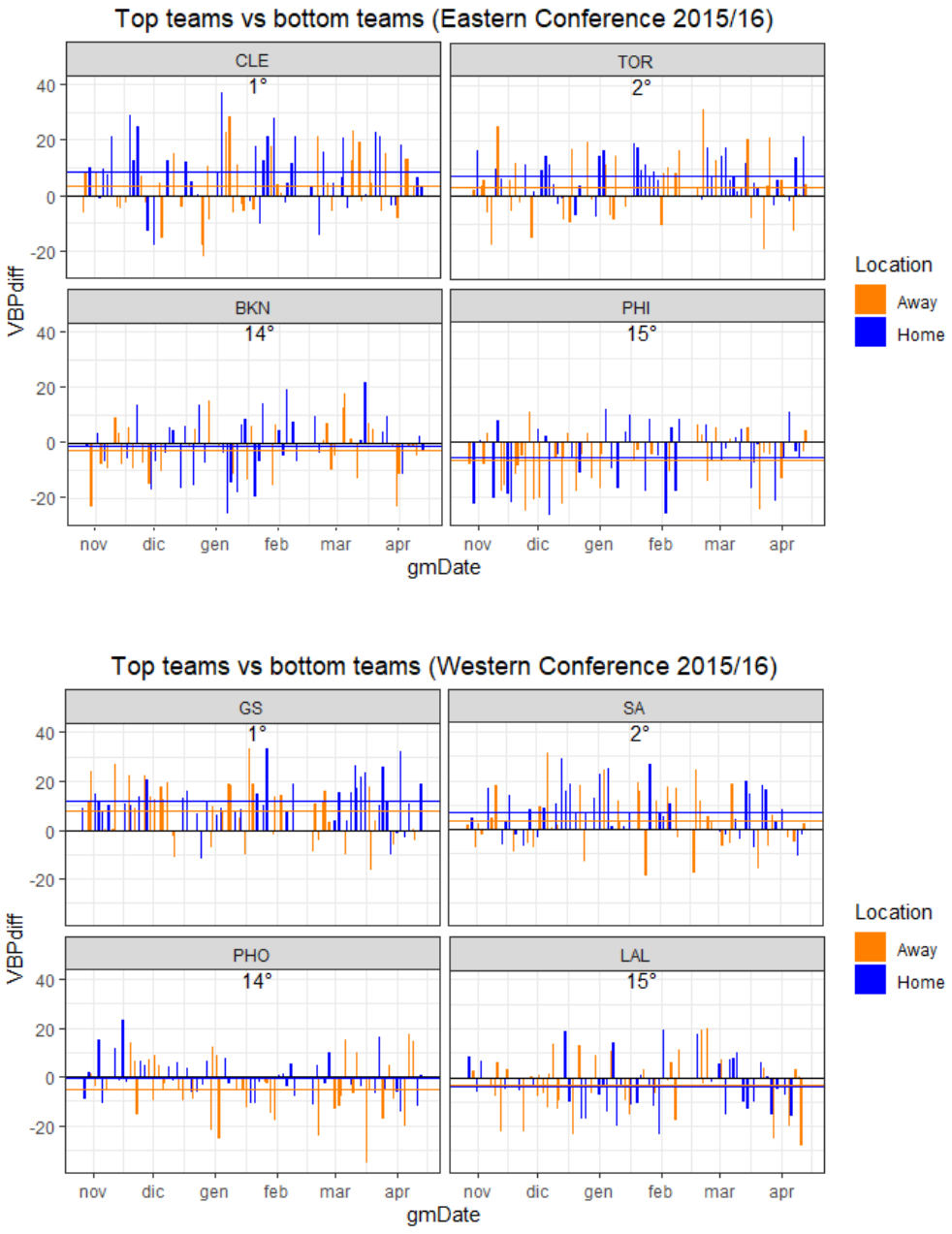


Figure 9 Home and away VBPdiff of the top and bottom Conference teams over 2015/16 season (9.a Eastern Conference, 9.b Western Conference)

2.3 – Predictive Analysis

The ultimate goal of the analysis is to predict the outcome of a game having at disposal all the variables introduced in the sections above. In the first part of this section, some techniques to transform the actual variables in usable ones for the predictive task are shown. Then, several models are developed for this purpose. Finally, a comparison between them is done, in order to understand which has the best predictive performance on new data and which predictors are the most significant to decree the winner of a game.

All the functions mentioned below come from the *caret* package (short for Classification and REgression Training) created and maintained by Max Kuhn. Moreover, due to extremely high computational time required for the training and tuning of several models, the computations are spread across five cores in parallel, thanks to the *doParallel* package[77].

2.3.1 - Data Preparation II

As stated in the last part of the Data Preparation I section, the data, even if they were cleaned and enriched with relevant information, are currently unusable for a predictive task. In fact, each row of the dataset contains details about the game after it was played, with the exception of few variables, such as the *Elo Ratings*, *DayOff* and the ones related to game details, like the date or the names of the teams on court, that are available prior to the start of the game. Hence, being that the last goal of the analysis is to predict the outcome of the games (win or loss), these data are useless, and so it is necessary to obtain all the available information before the start of the game.

[77] https://cran.r-project.org/web/packages/doParallel/index.html

Therefore, in the following two subsections, the adjustments required for extracting useful information are shown. The former explains how to deal with standings data and introduces two methods for getting recent teams' performances, while the latter exposes some common statistical techniques for removing redundant or not significative predictors.

2.3.1.1 - Getting Usable Data

Regarding to the variables that were originally included in the *standings* dataset (i.e., streaks, cumulated and average points scored, Pythagorean winning percentage and others) the operation needed to get usable data is straightforward. In fact, by simply shifting back to one game each of these (season by season), you can get predictors that can be used for the desired task. This means that, for the first game played by each team in a specific season, no information will be provided, but, starting from the second game, the data of the previous game will be displayed, and so on up to the last game of each season, which will include all details available after the second to last game played. Almost all the variables of this kind are therefore lagged in this way, but there are few exceptions: since that in the NBA at least one game is played almost every day, and above all the schedules can be very twisted due to several reasons (court availability, broadcasters, travel distance, etc.), *rank* and *gmBack* variables cannot be shifted back. The same issue arises for *opptW/L%* and *opptOpptW/L%* and for the related metrics already introduced, *SOS*, *RPI* and *SRS*. Thus, these variables (both for home and away team) are removed from the dataset. The new predictors replace the old ones, and their names differ for the suffix "*.bef*", standing for "before the start of the game".

For the remaining variables, this transformation would be meaningless. Hence, there is a need for finding a better solution for the variables originally included in the

BoxScores dataset. Intuitively, the best answer to the issue is to summarise the recent team's performance by computing the mean of the statistics and metrics recorded in their previous games. Two distinct methods to achieve this goal are introduced below: *Simple Moving Average* (*SMA*) and *Exponential Weighted Average* (*EWA*).

The former, as suggested by the name, works in a straightforward way: once a value for the parameter *k* is chosen arbitrarily, the average values of the statistics of interest are computed for the last *k* games played. The choice of the width of the moving window (*k*) is crucial: if it is too small, there is a high risk not to catch the performance's trend but instead the average value depends only on the very last games, on the contrary, if it is too big, the risk is "to smooth" too much the estimates, not considering the recent performance but instead approaching more to seasonal averages. Here the value of *k* is chosen equals to 7, to find a compromise between the two undesirable scenarios. Also, the new variables are later shifted back to one game, exactly as before, in such a way that the information become available before the start of the game. Hence, on the first game, no information is available, on the second one, the value of the first game is displayed, on the third one, the average between the first and the second games' values are considered, and so on up to the eighth game, when the value shown is the mean of the seven previous games' values. Starting from the ninth game, the window begins "to move", excluding the first game from the calculation of the average, till to the eighty-second game, where the value shown is the average of the previous seven games' values. The new variables' names differ from the old ones for the suffix "*.av*".

The latter, *EWA*, overcomes the problem of the arbitrary choice for the parameter *k*, by using an increasing version of the so-called "exponential smoothing". Basically, for each game, the statistics of interest are computed by taking the weighted average of all the previous games, with weights that increase exponentially. Once again, the new value

is then shifted back to one game in order to get the information available before the start of the game. Hence, on the first game, no information is available, on the second one, the value of the first game is shown, on the third one, the value displayed is the weighted average of the first two games' values, with the second game having a greater weight than the first one, and so on till the eighty-second game, where the value of the statistic of interest correspond to the weighted average of the eighty-one games previously played, with weights that start from nearly zero for the first game, and that exponentially increase till the second to last game, that has the greatest weight among the others. The new variables' names differ from the old ones for the suffix "*.ewa*". The exact formula for the calculation of the *EWA* for a sequence of *n* values is the following[78]:

$$EWA_\rho(x_1, \dots, x_n) = \frac{1-\rho}{1-\rho^n}(\rho^{n-1}x_1 + \rho^{n-2}x_2 + \cdots + x_n)$$

Where ρ is the smoother parameter that ranges from 0 to 1: as it grows, the *EWA* approaches to the arithmetic mean, on the contrary, if it is close to 0 the *EWA* approaches to the last value of the sequence. Here the value of ρ is set equal to 0.75.

The following plot (figure 10) shows the differences between the two techniques by taking as an example the Pace of Atlanta Hawks during 2012/13 season. The *SMA* (blue line) and the *EWA* (red line) seem having an analogous behaviour, even if the former looks smoother, while the latter appears to be more affected by the last games and thus, it has a more uneven shape.

78 https://stats.stackexchange.com/questions/286640/definition-of-the-function-for-exponentially-decaying-weighted-average

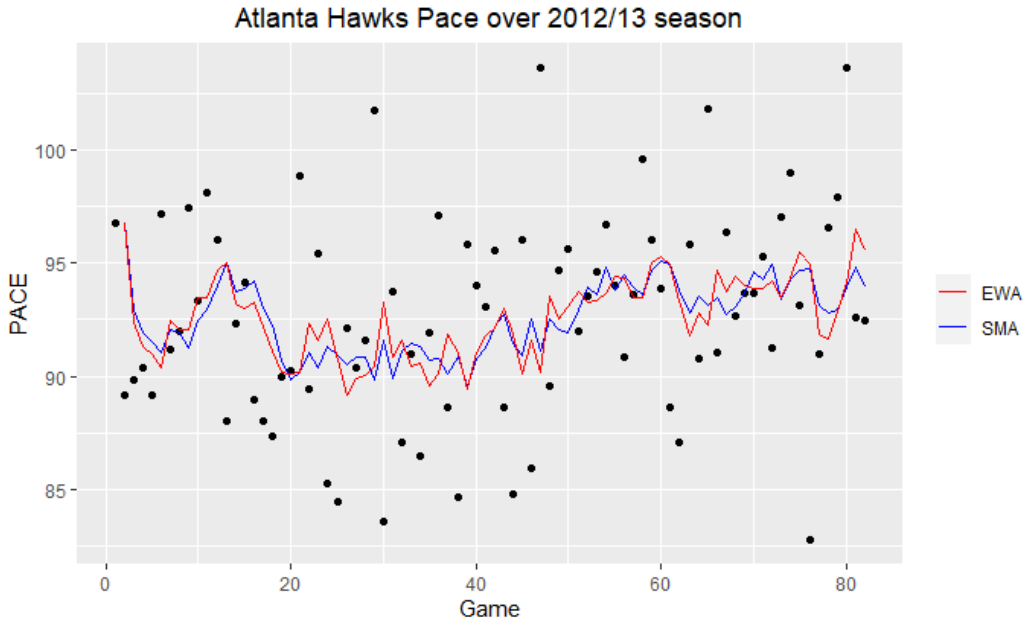


Figure 10. *Simple Moving Average* with k=7 (blue line) and *Exponential Weighted Average* with ρ=0.75 (red line) for the Atlanta Hawks *Pace* over 2012/13 season

Therefore, the old variables are removed from the dataset, and they are replaced with the new ones just computed, both with *SMA* and *EWA* methods.

Ultimately, a new variable in two different versions is computed: *log5*. *Matchup probabilities* or *log5* is a useful method invented by Bill James in 1981[79] for analysing baseball teams, but it can be easily applied to many other sports, like basketball for example. It is used for determining how often a team with a given winning percentage will beat another team with its winning percentage. In the formulations below, *log5* is based on *Pythagorean Winning Percentage* and also incorporates home court advantage, that is set to 0.6 (the league's home court teams win approximately 60% of the time). Two versions of *log5* are shown, the former relates to *Pythagorean Winning Percentage* with the 13.91 exponent before the start of the game, the latter to the 16.5 one. Clearly, *log5* is

79 James, Bill (1981) 1981 Baseball Abstract The 5th Annual Edition (1st edition)

calculated only for one team (here the home team, hence the "*H.*"), since that if the team's probability of winning the game equals to *P*, then their opponent's winning probability is *1-P*. The formula is the following[80]:

$$H.log5fpythx = \frac{H.pythx\%.bef \cdot (1 - A.pythx\%.bef) \cdot 0.6}{H.pythx\%.bef \cdot (1 - A.pythx\%.bef) \cdot 0.6 + (1 - H.pythx\%.bef) \cdot A.pythx\%.bef \cdot 0.4}$$

$$where\ x = 13.91\ or\ 16.5$$

Note that both James and Oliver formulation are based on the winning percentage in league, here the *Pythagorean* one is used trying to improve the method.

At the end of this phase the dataset is composed of 7379 rows and 311 columns.

2.3.1.2 - Pre-processing

Once the data have been adjusted making them usable for predicting the outcome of a game, the vary last step before moving to the Prediction phase is to remove the redundant and not significative predictors. This step is accomplished with four different sub-phases.

To begin, the rows with missing values are removed from the dataset. As stated above, they correspond to the first games played by teams on each season. The outcome of these games is in general arduous to predict due to the lack of information about teams at the very beginning of the season: in fact, even if teams tend to maintain the level of their performances from one season to the next, this is not always true. As shown in the *Elo Ratings* charts from 2.2.1 section, the performances can vary significatively across seasons and this is due to many reasons, including injuries and changes in roasters. Hence, each season has a certain degree of independence and thus, it is better to analyse them

80 http://www.rawbw.com/~deano/

separately. Just 103 of the 7379 rows feature missing values (less than 1.5%), therefore they are removed from the dataset.

The second step is related to *Zero-* and *Near-Zero Variance* predictors[81]. The former are those predictors that only have a single unique value (no predictor in the dataset has this feature), the latter are the ones having only a couple of unique values which occur with very low frequencies, and they may become *Zero-Variance* predictors when the data are split into sub-samples, causing errors in the computation of some models. The following rule of thumb identifies these predictors: if the fraction of unique values over the sample size is lower than 10% and the ratio of the frequency of the most common value to the frequency of the second most common value is larger than 19, then that variable is considered as a *Near-Zero Variance* predictor, and hence it is excluded from the analysis. There are 12 *Near-Zero Variance* predictors, all related to points scored during the overtime periods, that get delated from the dataset. This step was accomplished thanks to the *nearZeroVar* function.

The next step consists in detecting any eventual linear combination between the variables belonging to the dataset. Through a *QR decomposition*[82] of the data matrix, the sets of linear combinations get highlighted, and thus, the variables that can be removed to eliminate the linear dependencies are selected and delated. A total of 36 predictors are excluded from the analysis to counter the linear combinations. The *findLinearCombos* function was used for this purpose.

The last step regards pairwise correlation between predictors. Using highly correlated predictors in the development of a great number of models, can result in numerical errors, and low predictive performance. To deal with this issue, the solution is

81 Kuhn, Max, and Johnson, Kjell (2013) Applied Predictive Modeling, Springer, pp. 44-45
82 https://topepo.github.io/caret/pre-processing.html#linear-dependencies

to remove the minimum number of variables to guarantee that all pairwise correlations stay below a certain threshold. Thus, an intuitive algorithm is used for this purpose[83]:

1- Compute the correlation matrix of the predictors.
2- Determine the two predictors (named *A* and *B*) with the largest absolute pairwise correlation.
3- Compute the average correlation between *A* and the other predictors, same for *B*.
4- If *A* has a larger absolute average correlation, remove it; otherwise remove *B*.
5- Repeat steps 2-3-4 until no absolute correlations are above a given threshold.

By setting the threshold equal to 0.85, 161 highly correlated predictors are removed from the dataset thanks to the *findCorrelation* function.

All the methods and techniques introduced in this subsection relates to numerical predictors and cannot be applied to categorical variables, moreover, most models do not accept categorical predictors (except for tree-based models), unless they are encoded and converted to numerical values. The dataset contains a total of 14 categorical predictors, five of which with more than 50 levels, therefore encoding them would result in a dramatic increase in the number of predictors and ultimately in computational time. After several attempts, it emerged that, excluding categorical predictors from the analysis, computational time and complexity decreased and at the same time the models' predictive performance remained stable. Thus, all the categorical predictors were removed from the dataset.

The final dataset, set for the Predictive Analysis, consists of 7276 games and 88 variables (87 predictors and the target variable).

83 Kuhn, Max, and Johnson, Kjell (2013) Applied Predictive Modeling, Springer, pp. 45-47

2.3.2 - Models

Prior to the start of the development of models' phase, a few essential operations are required.

First of all, after having set a seed for reproducibility, the data are split in two parts: 85% of the rows (6185 games) are randomly chosen to create the *training set*, the remaining 15% (1091 games) instead, constitutes the *test set*. As the names suggest, the former will be used in the training phase to give life to several predictive models which parameters will be tuned, if possible, the latter will be used at a later time, for testing and comparing the predictive performances of the models on new data.

After that, it is necessary to select the resampling type: here a *10-fold repeated-3-times cross-validation* is chosen for the analysis. Thus, during the training phase, the *training set* is randomly partitioned in 10 sets of roughly equal size called *folds*, and a model is fit using all the folds expect for the first one, that is used as *validation set* to estimates the predictive performance of the model. Then, the model is fit again using all the folds except for the second, which is used to estimates performance measures, and so on till all the *folds* have been used one time as *validation set*. The 10 resampled estimates are then summarized with a simple mean that is used to understand the model performance. Finally, the process is repeated 3 times and the average of these three values is considered, in order to increase precision while maintaining a small bias, as research has shown[84].

Ultimately, even if the target class' frequency is moderately balanced (58.5% of the results are wins, 41.5% are losses), class weights inversely proportional to their respective frequencies are computed and will be used in the development of some models

84 Kim, Ji-Hyun (2009) "Estimating Classification Error Rate: Repeated Cross–Validation, Repeated Hold–Out and Bootstrap." *Computational Statistics & Data Analysis*, Vol. 53, Iss. 11, pp. 3735–3745

(marked with "*case weights*" caption). Moreover, several models will be introduced in two different versions: the former uses all the predictors available for the fitting, the latter instead, uses only the predictors with a *VIF* smaller than 10. The *Variance Inflation Factor* (*VIF*) measures the multicollinearity between predictors in a regression model and is computed for each of them. Is defined as the ratio of the overall model variance (hence, the model with all predictors) to the variance of the model that includes only that predictor[85]. Thus, selecting predictors with a small *VIF*, will fix in part the issue of multicollinearity.

2.3.2.1 - Model Training

The full list of the fifty-one models used, arranged by family, is given below. For each of them the *caret* model's name is shown in brackets along with the final values of the tuning parameters if they exist. Each model has been processed prior to the training phase by *centring* and *scaling* the data (i.e. each variable has mean 0 and standard deviation 1), except for the ones marked with an asterisk. The *Area Under the ROC Curve* (*AUC*) is used as criterion to select the best tuning parameters. Each model was built with the *train* function. All the charts related to the *tuning* phase are displayed in the Appendix A.

- Generalized linear models:
    - Logistic Regression (*glm*)
    - Logistic Regression with case weights (*glm*)
    - Logistic Regression with *VIF<10* predictors (*glm*)
    - Logistic Regression with *VIF<10* predictors and case weights (*glm*)

85 James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2017) An Introduction to Statistical Learning (8th edition), Springer

- Bayesian Logistic Regression (*bayesglm*)
- Bayesian Logistic Regression with case weights (*bayesglm*)
- Bayesian Logistic Regression with *VIF<10* predictors (*bayesglm*)
- Bayesian Logistic Regression with *VIF<10* predictors and case weights (*bayesglm*)
- Boosted Logistic Regression (*glmboost*)

  Final tuning parameters:

  - *prune=”no”*
  - *mstop=400*
- Boosted Logistic Regression with case weights (*glmboost*)

  Final tuning parameters:

  - *prune=”no”*
  - *mstop=400*
- Boosted Logistic Regression with *VIF<10* predictors (*glmboost*)

  Final tuning parameters:

  - *prune=”no”*
  - *mstop=350*
- Boosted Logistic Regression with *VIF<10* predictors and case weights (*glmboost*)

  Final tuning parameters:

  - *prune=”no”*
  - *mstop=350*
- Partial Least Squares Regression (*pls*)

  Final tuning parameter:

  - *ncomp=8*

- Partial Least Squares Regression with *VIF<10* predictors (*pls*)

  Final tuning parameter:

  - *ncomp=9*

- Glmnet (*glmnet*)

  Final tuning parameters:

  - *alpha=0.94*
  - *lambda=0.0105*

- Discriminant Analysis models*:
  - Linear Discriminant Analysis (*lda*)
  - Linear Discriminant Analysis with *VIF<10* predictors (*lda*)
  - Penalized Discriminant Analysis (*pda*)

    Final tuning parameter:

    - *lambda=171000*

  - Penalized Discriminant Analysis with case weights (*pda*)

    Final tuning parameter:

    - *lambda=1947*

  - Penalized Discriminant Analysis with *VIF<10* predictors (*pda*)

    Final tuning parameter:

    - *lambda=171000*

  - Penalized Discriminant Analysis with *VIF<10* predictors and case weights (*pda*)

    Final tuning parameter:

    - *lambda=1947*

  - Regularized Discriminant Analysis (*rda*)

    Final tuning parameters:

        - *gamma=0.09*
        - *lambda=0.88*
    - Regularized Discriminant Analysis with *VIF<10* predictors (*rda*)

      Final tuning parameters:

        - *gamma=0.0009*
        - *lambda=1.1*

- Non-linear models:
    - K-Nearest Neighbours (*knn*)

      Final tuning parameter:

        - *k=349*
    - K-Nearest Neighbours with *VIF<10* predictors (*knn*)

      Final tuning parameter:

        - *k=416*
    - K-Nearest Neighbours with Principal Components Analysis (*knn*)

      Final tuning parameter:

        - *k=419*
    - K-Nearest Neighbours with *VIF<10* predictors and Principal Components Analysis (*knn*)

      Final tuning parameter:

        - *k=339*
    - Multivariate Adaptative Regression Splines (*earth*)

      Final tuning parameters:

        - *nprune=4*
        - *degree=1*

- Multivariate Adaptative Regression Splines with *VIF<10* predictors (*earth*)

  Final tuning parameters:

  - *nprune=4*
  - *degree=1*

- Neural Network with *VIF<10* predictors (*nnet*)

  Final tuning parameters:

  - *size=1*
  - *dacay=0.78*

- Neural Network with *VIF<10* predictors and case weights (*nnet*)

  Final tuning parameters:

  - *size=1*
  - *decay=0.24*

- Tree/Rules-based methods*:
  - CART (*rpart*)

    Final tuning parameter:

    - *cp=0.007*

  - CART with case weights (*rpart*)

    Final tuning parameter:

    - *cp=0.003*

  - Bagged CART (*treebag*)
  - Bagged CART with case weights (*treebag*)
  - Conditional Inference Tree (*ctree*)

    Final tuning parameter:

    - *mincriterion=0.94*

- Conditional Inference Tree with case weights (*ctree*)

  Final tuning parameter:

  - *mincriterion=0.99996*

- Boosted Tree (*blackboost*)

  Final tuning parameters:

  - *mstop=50*
  - *maxdepth=3*

- Boosted Tree with case weights (*blackboost*)

  Final tuning parameters:

  - *mstop=50*
  - *maxdepth=3*

- C5.0 (*C5.0*)

  Final tuning parameters:

  - *trials=20*
  - *model=”rules”*
  - *winnow=FALSE*

- C5.0 with case weights (*C5.0*)

  Final tuning parameters:

  - *trials=13*
  - *model=”rules”*
  - *winnow=FALSE*

- Random Forest (*ranger*)

  Final tuning parameters:

  - *mtry=34*
  - *splitrule=”extratrees”*

        - *min.node.size=225*

- Random Forest with case weights (*ranger*)

    Final tuning parameters:

        - *mtry=34*
        - *splitrule=”extratrees”*
        - *min.node.size=275*

- eXtreme Gradient Boosting (*xgbTree*)

    Final tuning parameters:

        - *nrounds=350*
        - *max_depth=2*
        - *eta=0.025*
        - *gamma=11*
        - *colsample_bytree=0.4*
        - *min_child_weight=250*
        - *subsample=1*

- eXtreme Gradient Boosting with case weights (*xgbTree*)

    Final tuning parameters:

        - *nrounds=250*
        - *max_depth=2*
        - *eta=0.025*
        - *gamma=1*
        - *colsample_bytree=0.4*
        - *min_child_weight=15*
        - *subsample=1*

- Support Vector Machines:
    - Support Vector Machines with Linear Kernel (*svmLinear*)

      Final tuning parameter:

        - *C=0.0058*
    - Support Vector Machines with Linear Kernel with *VIF<10* predictors (*svmLinear*)

      Final tuning parameter:

        - *C=0.0068*
    - Support Vector Machines with Radial Basis Function Kernel (*svmRadial*)

      Final tuning parameters:

        - *Sigma=0.001*
        - *C=1*
    - Support Vector Machines with Radial Basis Function Kernel with *VIF<10* predictors (*svmRadial*)

      Final tuning parameters:

        - *Sigma=0.001*
        - *C=2*
    - Support Vector Machines with Polynomial Kernel (*svmPoly*)

      Final tuning parameters:

        - *degree=5*
        - *scale= 0.00001*
        - *C=100*
    - Support Vector Machines with Polynomial Kernel with *VIF<10* predictors (*svmPoly*)

      Final tuning parameters:

- *degree=4*
- *scale= 0.00001*
- *C=175*

2.3.2.2 - Model Comparison and Results

Before to proceed with the comparison of models' predictive performances, the analysis focuses on the *variable importance plot* of each model. *Variable importance*[86] refers to how much a certain model "uses" a specific variable to produce accurate predictions. The more the model relies on a variable to make predictions, the more "important" it is. The results of the analysis are shown in the table below, achieved thanks to the *varImp* function: the top five most important predictors for each model are depicted from left to the right, the variable's name along with the number of models that selected it are displayed. Out of the fifty-one models, thirty-eight selected the *log5 Pythagorean Winning Percentage built with the 13.91 exponent* as first, twelve the *home team Elo Rating*, and only one model (*C5.0 with case weights*) the *number of lost games in Conference by the away team*. Regarding the second predictor in order of importance, thirty-seven models chose the *home team Elo Rating*, thirteen the *away team Elo Rating*, and only one model (*Glmnet*) the *exponential weighted average of the away team percentage of field goal made by assist*. Respecting to the third, fourth and fifth variables instead, the most "used" ones are the *home team winning percentage*, the *away team Elo Rating*, and the *away team winning percentage*, respectively. Therefore, is clear that some variables are more important than others to produce accurate predictions. Specifically, three of them stand out among the others: *log5*, *home team Elo Rating* and *away team Elo Rating*, which are not selected among the top five predictors only by two models. Other

86 https://stats.stackexchange.com/questions/332960/what-is-variable-importance

important variables in a predictive perspective seem the *home and away winning percentage* and the *number of days since last game played by away team*. Moreover, more than half (twenty-eight) of the models "use" the following variables in this exact order: *H.log5fpyth1391*, *H.elobefore*, *H.W/L%.bef*, *A.elobefore*, *A.W/L%.bef*. The majority of the models seem preferring the same variables to make accurate prediction, but one of them stand out among the others for some apparently foolish choices. In fact, the *Glmnet* top five variables are in the order: *H.log5fpyth1391*, *A.FGMAST%.ewa*, *H.STL.ewa*, *A.PTS6.ewa*, *A.PTS1.av*. Nevertheless, this model will turn out to have a good predictive performance, meaning that it "works" in a unique manner among the others. All the *variable importance plot* can be found in the Appendix B.

Variable 1	**Variable 2**	**Variable 3**	**Variable 4**	**Variable 5**
H.log5fpyth1391(38) H.elobefore(12) A.confL.bef(1)	H.elobefore(37) A.elobefore(13) A.FGMAST%.ewa(1)	H.W/L%.bef(30) H.log5fpyth1391(9) A.elobefore(6) A.DayOff(2) H.PTS1.av(2) H.lastFive.bef(1) H.STL.ewa(1)	A.elobefore(30) H.W/L%.bef(6) A.PTS1.av(4) A.2PA.av(2) A.DayOff(2) Other(7)	A.W/L%.bef(33) A.DayOff(3) A.PTS1.av(3) A.3P%.ewa(2) A.FG%.av(2) Other(8)

In order to compare the predictive performances of the models on new data (*test set*), the *confusion matrix* of each of them was computed thanks to the *confusionMatrix* function. Figure 11 displays the result of the comparison. The black dot is the *accuracy* value of each model, while the light blue line represents the 95% confidence interval. Starting from the bottom of the plot, the *Conditional Inference Tree* is the one with the lowest *accuracy* (0.63978). The simple *logistic regression* (0.6672777), usually considered as a benchmark, performs better than all the *K-Nearest Neighbours* models

and few others. By moving upwards, the first model that overcomes the value of 0.68 is the *Partial Least Squares regression* (0.68011). The top five models according to the analysis are, in the order: *Penalized Discriminant Analysis* (0.6865261), *Boosted Logistic Regression* (0.6856095), *Bayesian Logistic Regression* (0.6856095), *Random Forest* (0.6846929) and *Linear Discriminant Analysis* (0.6846929). The *confusion matrix* of the *Penalized Discriminant Analysis* is shown in figure 12.a: the model correctly predicted the outcome of 749 games out of the 1091 available ones, even if slightly above the 50% of the actual losses (231 out of 452) were properly predicted. Figure 12.b instead, is related to the *Boosted Logistic Regression*: this model correctly predicted only one game less than the previous one (748), but two more losses (233).

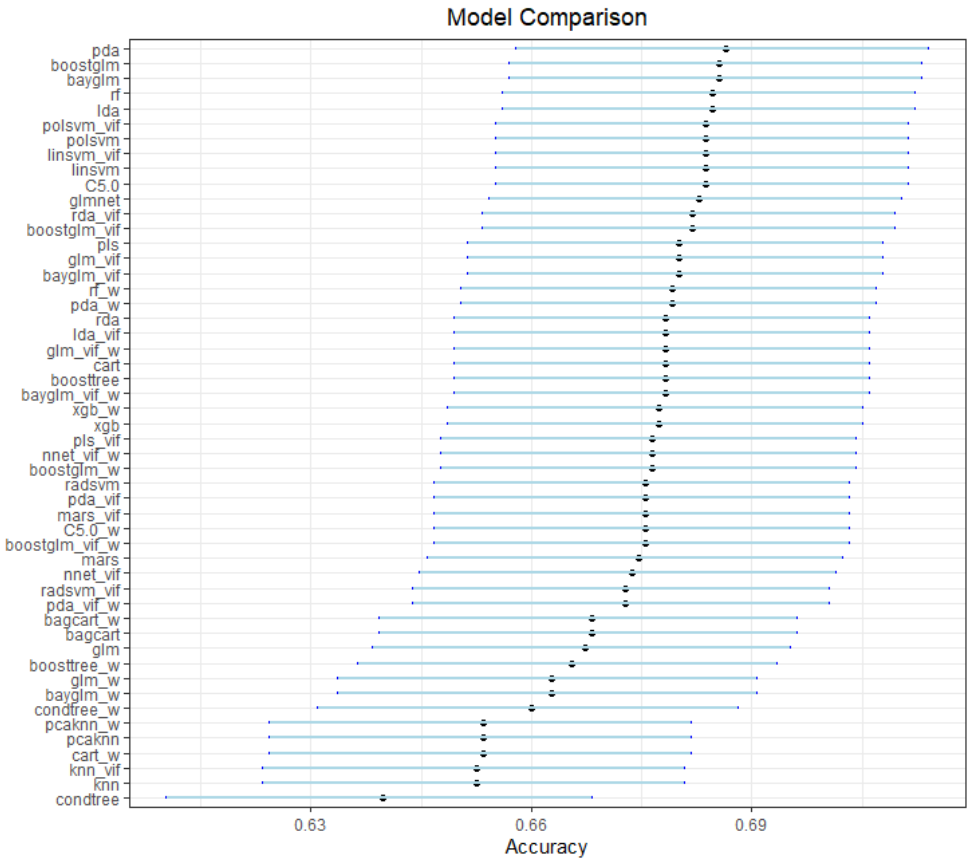


Figure 11. Comparison of the models' *accuracy* on *test set*, the light blue line stands for the 95% confidence interval

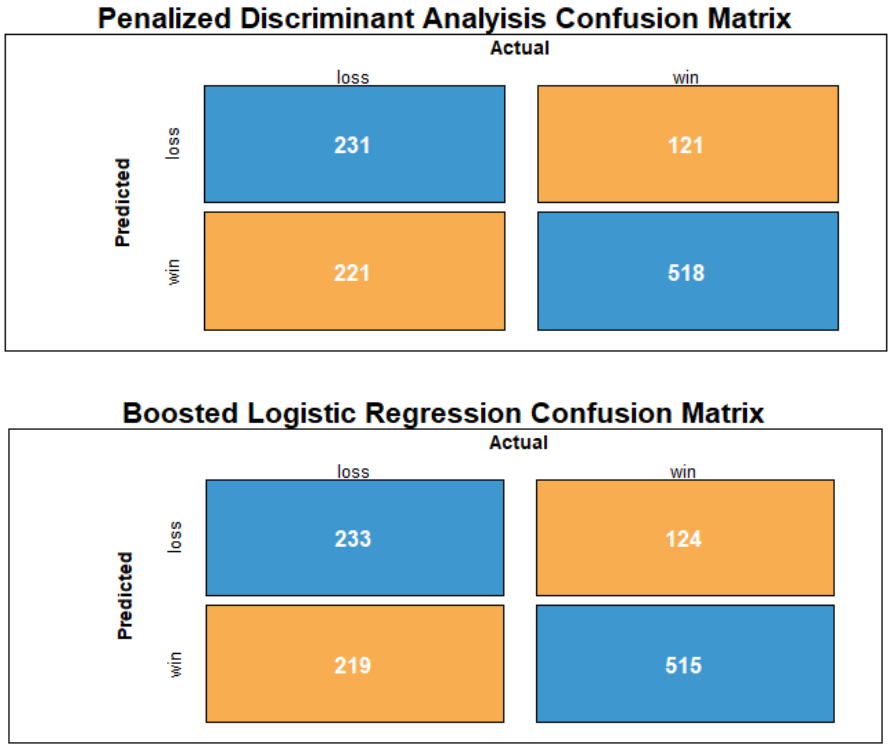


Figure 12. Analysis of the *confusion matrix* on *test set* of the *Penalized Discriminant Analysis* (figure 12.a) and of the *Boosted Logistic Regression* (figure 12.b)

Ultimately, trying to improve the *overall accuracy* of the predictions and the number of losses properly predicted, two basic *ensemble* techniques were adopted: *simple averaging* and *majority voting*. The former combines the predicted probabilities of two or more models by taking the simple mean, the latter instead, considers the mode of the class labels predicted by two or more models. The goal of ensembling methods is to combine the predictions of several base estimators in order to improve generalizability / robustness over a single estimator[87]. Here this goal is accomplished by combining three models: *Penalized Discriminant Analysis, Boosted Logistic Regression with VIF<10 predictors and case weights* and *C5.0*. The *confusion matrix* related to the former method (figure 13.a) highlight a clear improvement in both desired aims: the ensemble model correctly predicted 757 game results, producing an *overall accuracy* equals to 0.6939,

[87] https://scikit-learn.org/stable/modules/ensemble.html#

and 271 losses properly predicted. The model built with the latter method (figure 13.b) achieved similar results: although the *overall accuracy* is slightly lower than the previous technique (0.6911, with 754 game outcomes correctly predicted), the number of losses properly predicted was 282, more than the 62% of the actual losses.

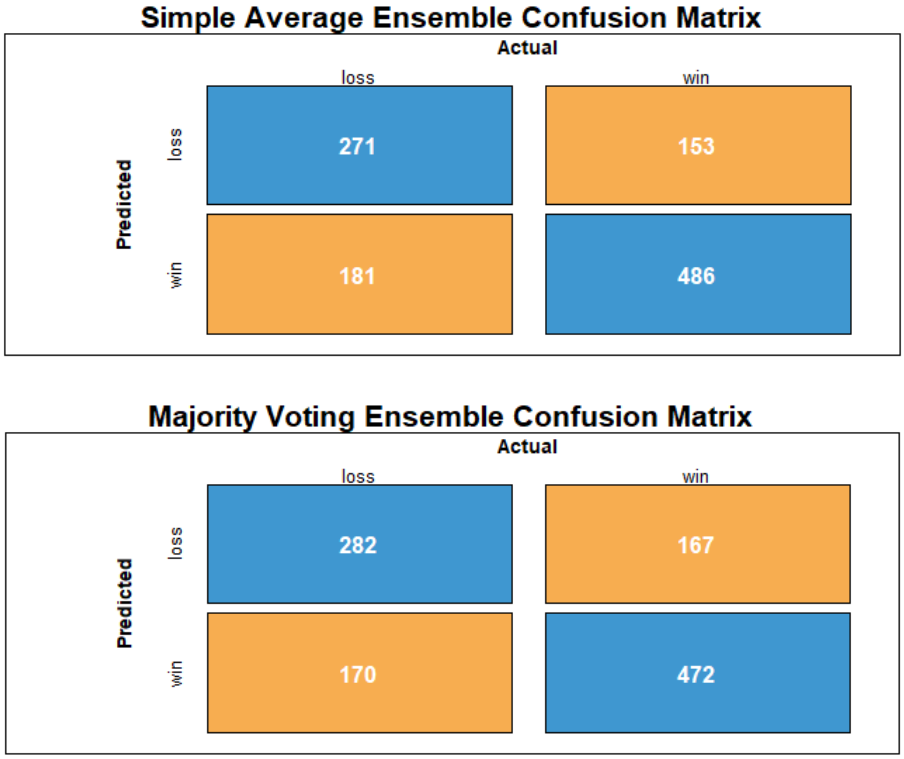


Figure 13. Analysis of the *confusion matrix* on *test set* of ensemble model built with *Penalized Discriminant Analysis, Boosted Logistic Regression with VIF<10 predictors and case weights* and *C5.0* (figure 13.a Simple Average, figure 13.b Majority voting)

2.4 Conclusions

This research aimed to explore basketball data by using the traditional statistical tools. In the first part of the analysis was revealed that the Eastern and the Western Conference are not on the same average level of performance. By visualizing the evolution of the *Elo Ratings* across the analysed seasons, the western teams turned out to have a superior performance on average, and regarding to Divisions, the Pacific and the Southwest one distinguished among the others for being an authentic producer of the so-called "super teams". Then, switching the focus on *VBPdiff*, this scenario of disparity is confirmed, and the seasonal averages of the new statistic revealed that having a considerably high (low) value of *VBPdiff* translates into a top (bottom) ranking at the end of the regular season.

The second part of the analysis aims to predict the outcome of regular season games by using as predictors the recent team's performance collected in the previously played games. The comparison of more than fifty predictive models showed that not all the variables have the same importance in producing accurate predictions. It turned out that the already mentioned *Elo Ratings* and the *log5 built with Pythagorean Winning Percentage* instead of the classical winning percentage, stand out among the others in this specific task. The final model, an ensemble of *Penalized Discriminant Analysis*, *Boosted Logistic Regression* and *C5.0*, correctly predicted the 69.39% of the game in the *test set*. This result, although it may be improved by collecting more data and having available more computational time to spend on the training of models, should be

considered satisfactory, since most of the previous research achieved an *accuracy* that ranges from 64% to 71%[88] [89] [90] [91] [92] [93].

Basketball analytics represents a really interesting and challenging field: a basketball game may produce many kinds of data type and the modern technologies are able to record and collect each of them. Here, the basic one was analysed, but, for future works, the use of some more advanced data types could improve the predictive performances. This research concludes hoping that soon, by the improvement of the actual metrics and rating systems and by the development of new ones, there will be a significant increase in *accuracy* of predictions.

[88] Lieder, Nachi (2018) Can Machine-Learning Methods Predict the Outcome of an NBA Game? p. 9
[89] Liu, Jizhi (2020) Predicting United States National Basketball Games using Machine Learning Techniques
[90] Puranmalka, Keshav (2013) Modelling the NBA to make better predictions
[91] http://dionny.github.io/NBAPredictions/website/
[92] Uudmae, Jaak (2017) Predicting NBA Game Outcomes p.1
[93] Prastuti, Singh; Bai Wang, Yang (2019) NBA Game Predictions based on Player Chemistry p. 4

Bibliography

James, Bill (1981) 1981 Baseball Abstract The 5th Annual Edition (1st edition)

Elo, Arpad (1986) The Rating of Chessplayers, Past and Present (2nd edition), Arco

Dewan, John; Zminda, Don (1993) STATS 1993 Basketball Scoreboard (1st edition), Harperreference

Imbrogno, Raffaele (2004) Statistica e Pallacanestro

Kubatko, Justin; Oliver, Dean; Pelton, Kevin; and Rosenbaum, Dan T. (2007) "A Starting Point for Analyzing Basketball Statistics," *Journal of Quantitative Analysis in Sports*: Vol. 3: Iss. 3, Article 1

Kim, Ji-Hyun (2009) "Estimating Classification Error Rate: Repeated Cross–Validation, Repeated Hold–Out and Bootstrap." *Computational Statistics & Data Analysis*, Vol. 53, Iss. 11

Puranmalka, Keshav (2013) Modelling the NBA to make better predictions (https://dspace.mit.edu/handle/1721.1/85464)

Kuhn, Max, and Johnson, Kjell (2013) Applied Predictive Modeling (1st edition), Springer

Uudmae, Jaak (2017) Predicting NBA Game Outcomes (http://cs229.stanford.edu/proj2017/final-reports/5231214.pdf)

James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2017) An Introduction to Statistical Learning (8th edition), Springer

Lieder, Nachi (2018) Can Machine-Learning Methods Predict the Outcome of an NBA Game? (https://ssrn.com/abstract=3208101)

Prastuti, Singh; Bai, Yang (2019) NBA Game Predictions based on Player Chemistry (http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26645648.pdf)

Liu, Jizhi (2020) Predicting United States National Basketball Games using Machine Learning Techniques (https://ssrn.com/abstract=3827474)

Zuccolotto, Paola; Manisera, Marica (2020), Basketball Data Science – with Applications in R. Chapman and Hall/CRC

Chang, Wesley; Ran, Michael; Smith, Gary (2021) The Impacts of Home-Court Advantage in the NBA (http://economics-files.pomona.edu/GarySmith/Econ190/Econ190%202021/ChangRan.pdf)

Sitography

http://www.adsoftheworld.com

http://www.basketball.realgm.com

http://www.basketballforcoaches.com

http://www.basketball-references.com

http://www.basketcaffe.com

http://www.biography.com

http://www.ca.nba.com

http://www.careers.nba.com

https://www.dplyr.tidyverse.org/

http://www.espn.com

http://www.fivethirtyeight.com

http://www.forbes.com

http://ggplot2.tidyverse.org/

http://www.github.com

http://www.gleague.nba.com

http://www.hooptactics.com

http://www.kaggle.com

http://www.mapsontheweb.com

http://www.mlb.com

http://www.nationalgeographic.com

http://www.nba.com

http://www.nbahoopsonline.com

http://www.nbastuffer.com

http://www.official.nba.com

http://www.olympedia.com

http://www.pr.nba.com

http://www.rarenewspaper.com

http://www.rawbw.com/~deano/

http://www.r-project.org

http://www.rstudio.com

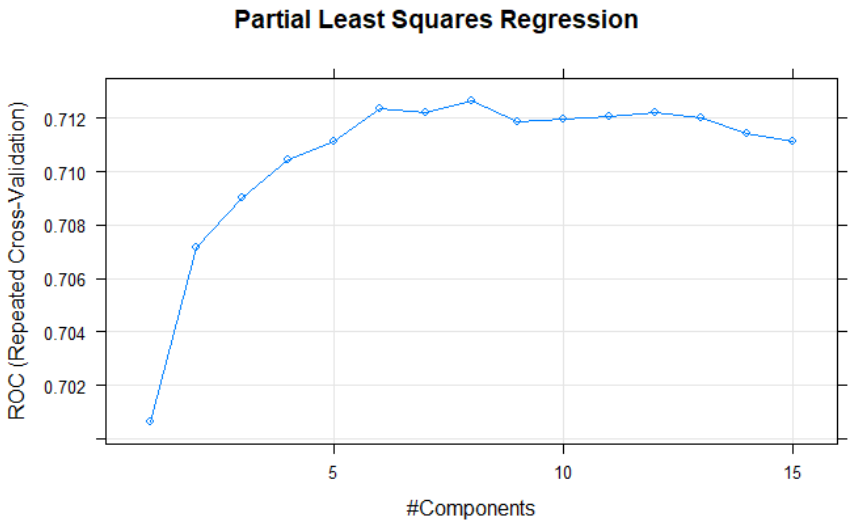
http://www.scikit-learn.org

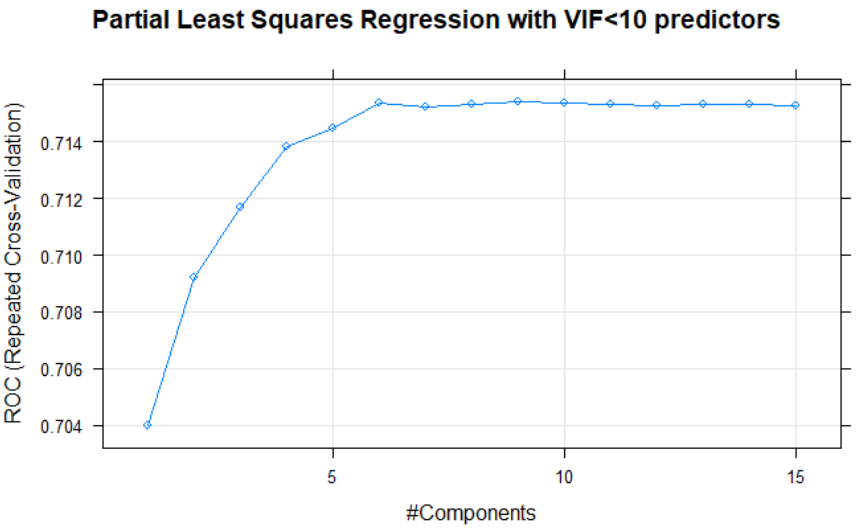
http://www.sonicrising.com

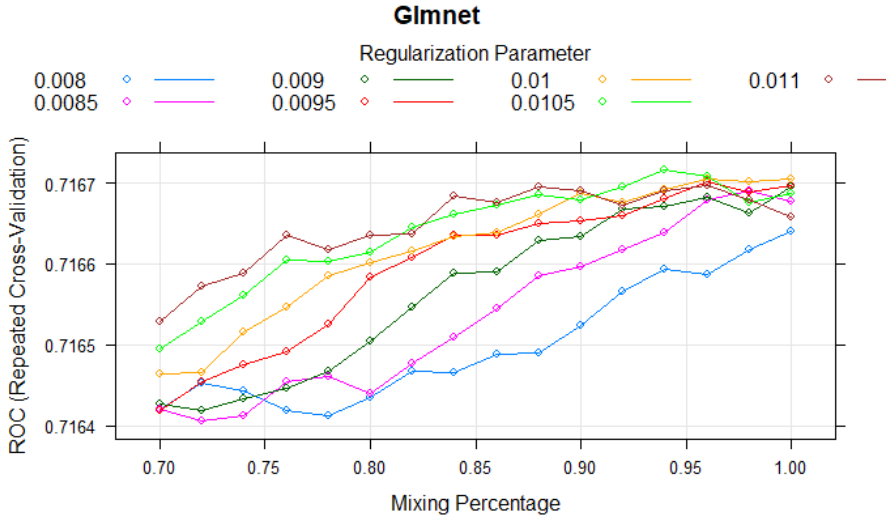
http://www.statista.com
http://www.stats.stackexchange.com
http://www.topepo.github.com
http://www.usab.com
http://www.web.archive.org
http://www.webcitation.org

Appendix A

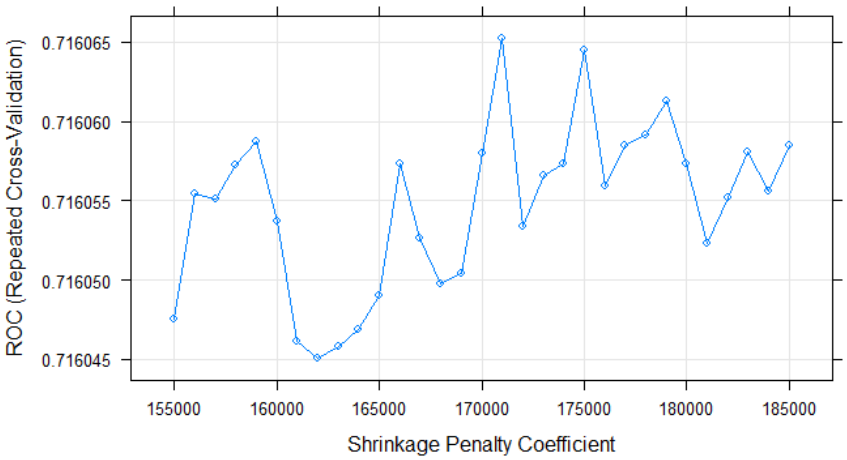
Tuning Parameters



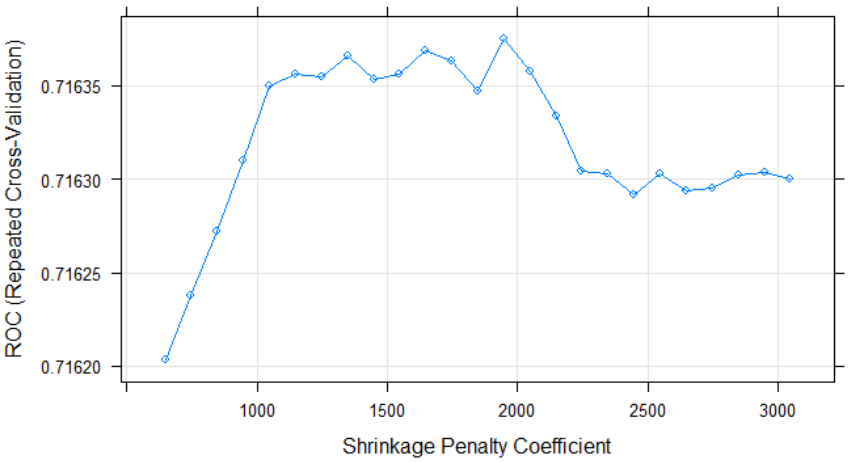




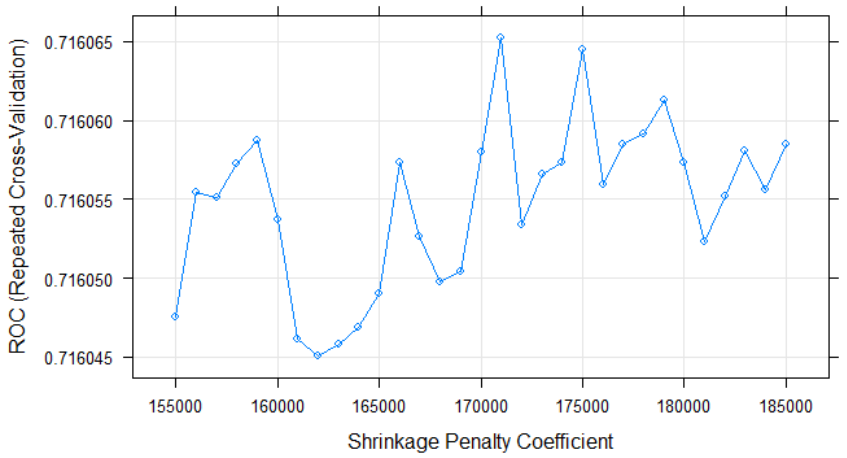
Penalized Discriminant Analysis



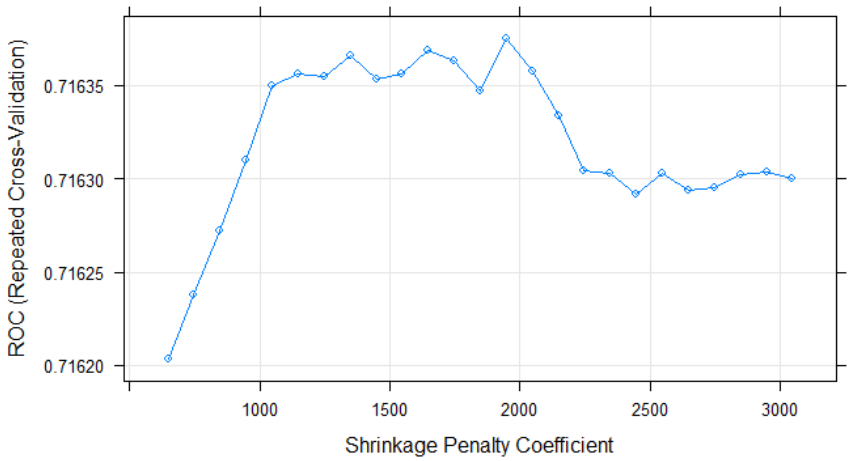
Penalized Discriminant Analysis with VIF<10 predictors



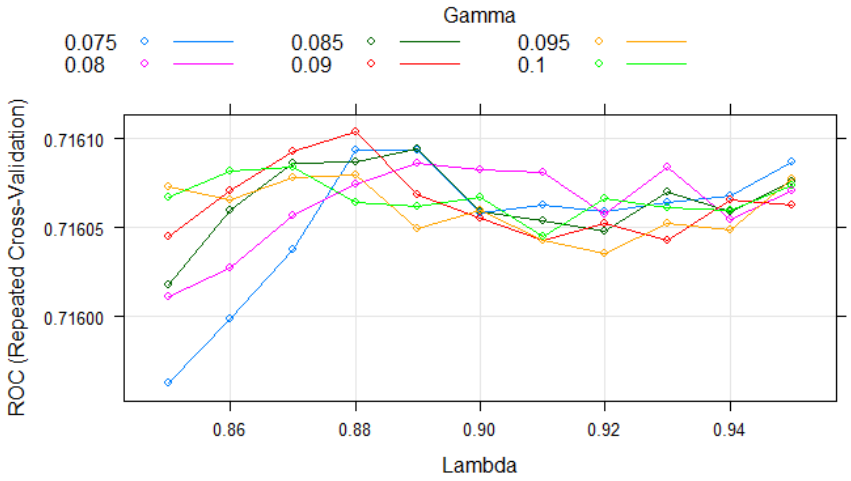
Penalized Discriminant Analysis with case weights



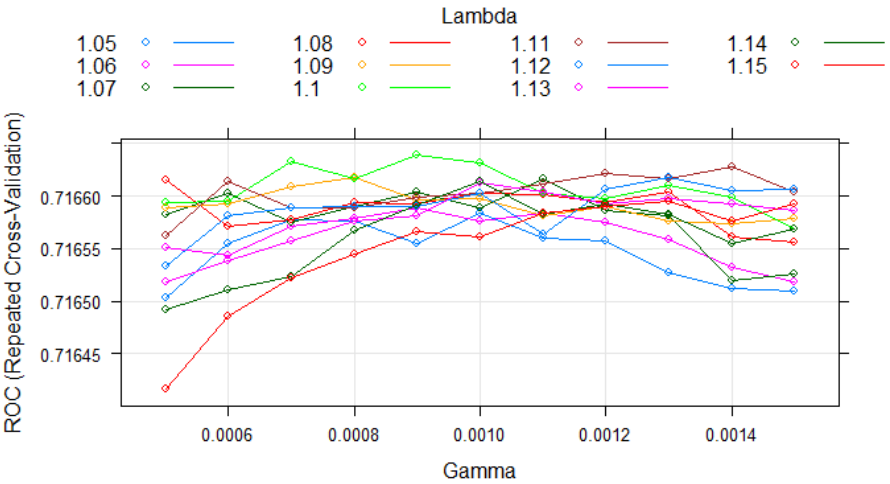
Penalized Discriminant Analysis with VIF<10 predictors and case weights

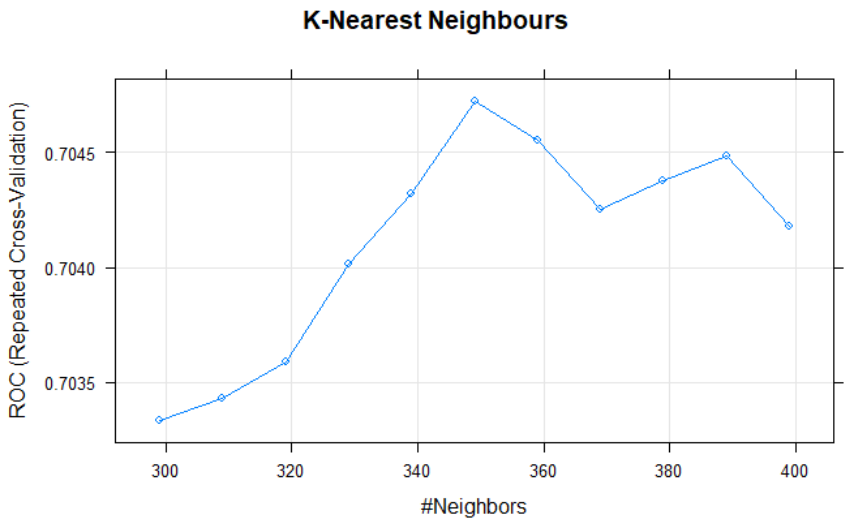


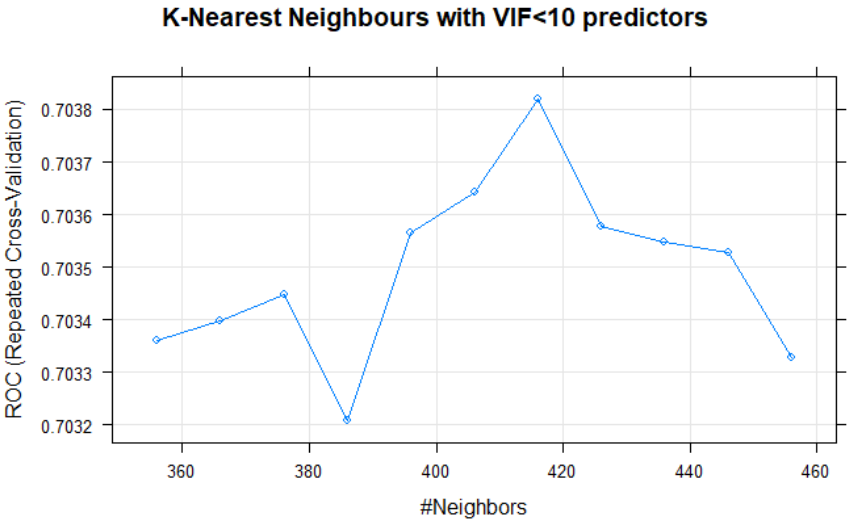
Regularized Discriminant Analysis

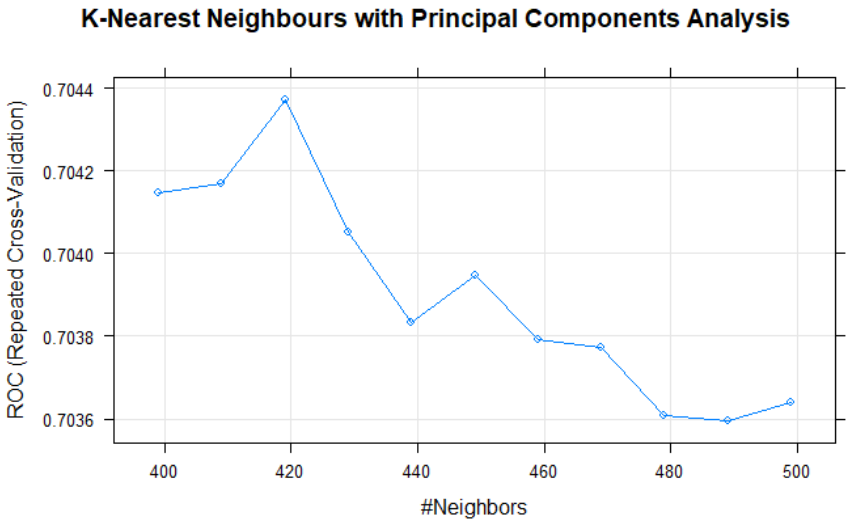


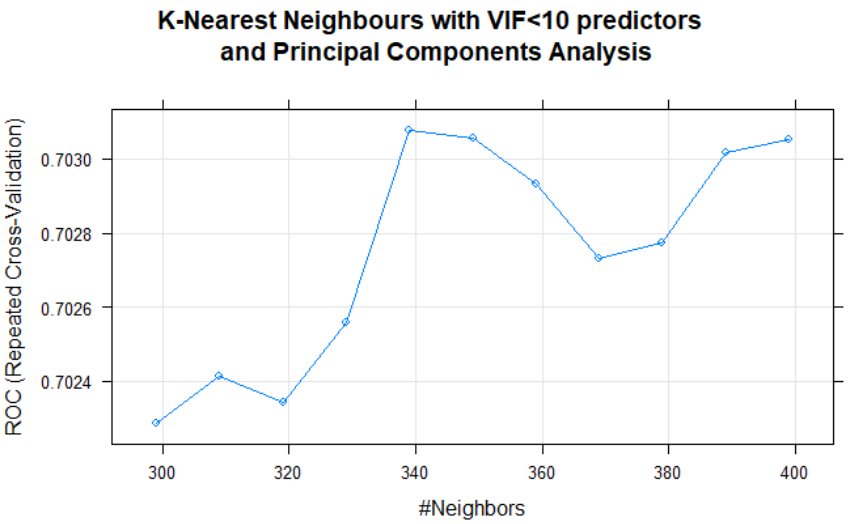
Regularized Discriminant Analysis with VIF<10 predictors

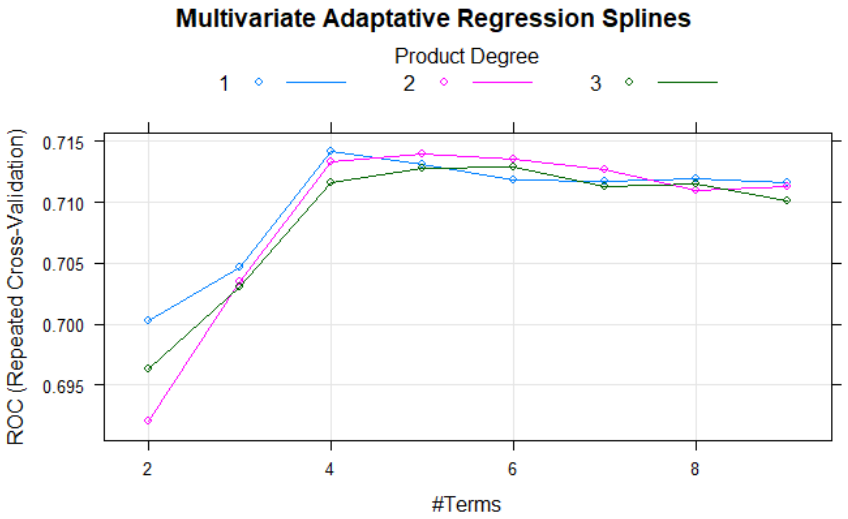


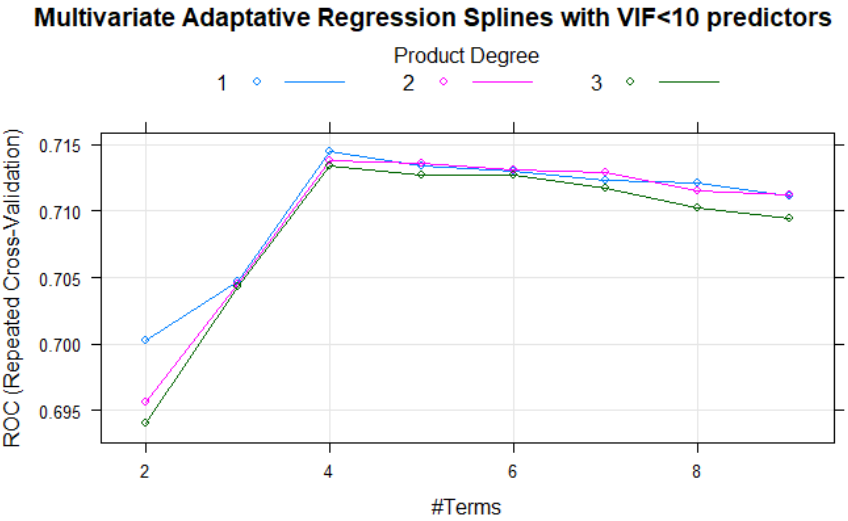

K-Nearest Neighbours
ROC (Repeated Cross-Validation)
0.7045
0.7040
0.7035
300
320
340
360
380
400
#Neighbors

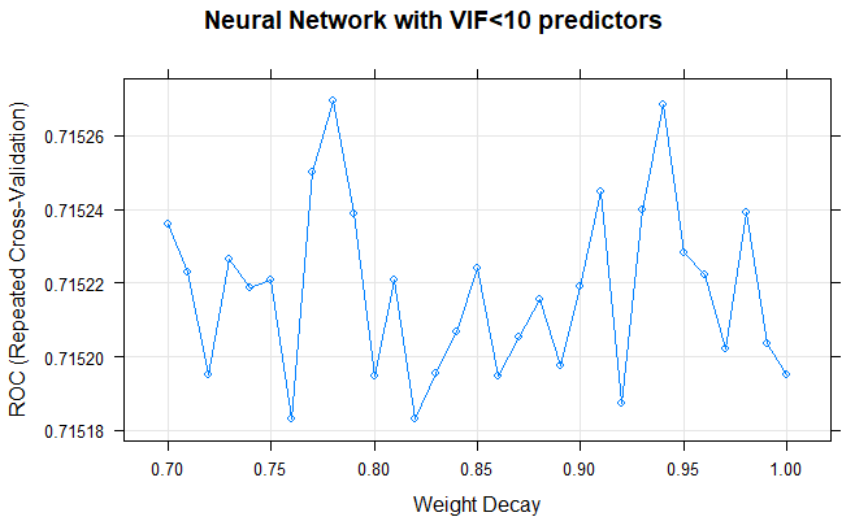

K-Nearest Neighbours with VIF<10 predictors
ROC (Repeated Cross-Validation)
0.7038
0.7037
0.7036
0.7035
0.7034
0.7033
0.7032
360
380
400
420
440
460
#Neighbors

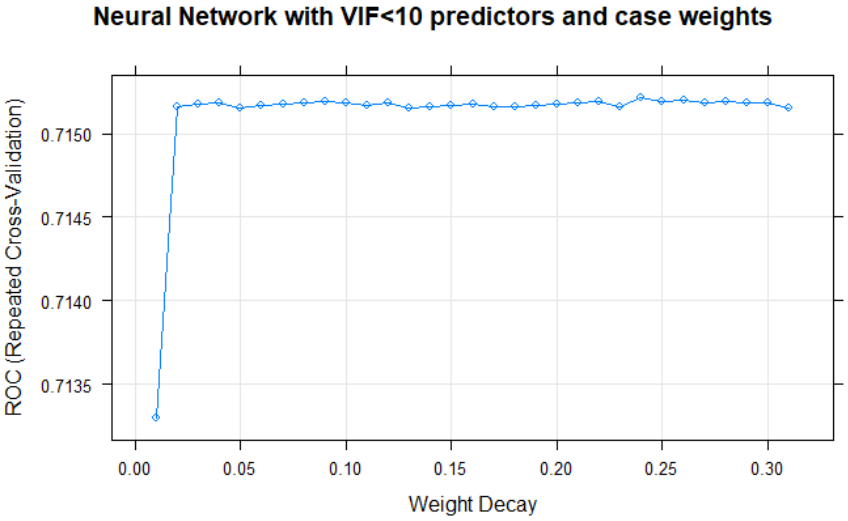

K-Nearest Neighbours with Principal Components Analysis
ROC (Repeated Cross-Validation)
0.7044
0.7042
0.7040
0.7038
0.7036
400
420
440
460
480
500
#Neighbors

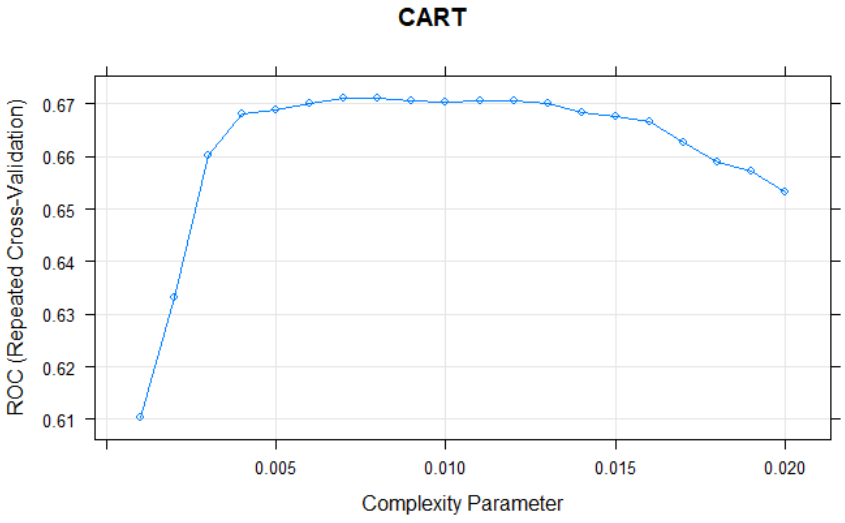

K-Nearest Neighbours with VIF<10 predictors
and Principal Components Analysis
ROC (Repeated Cross-Validation)
0.7030
0.7028
0.7026
0.7024
300
320
340
360
380
400
#Neighbors

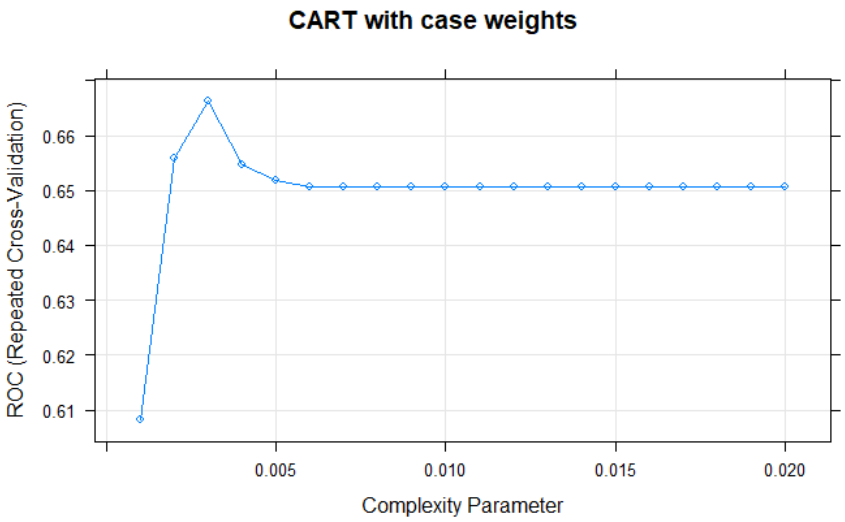

Multivariate Adaptative Regression Splines
Product Degree
1
2
3
ROC (Repeated Cross-Validation)
0.715
0.710
0.705
0.700
0.695
2
4
6
8
#Terms

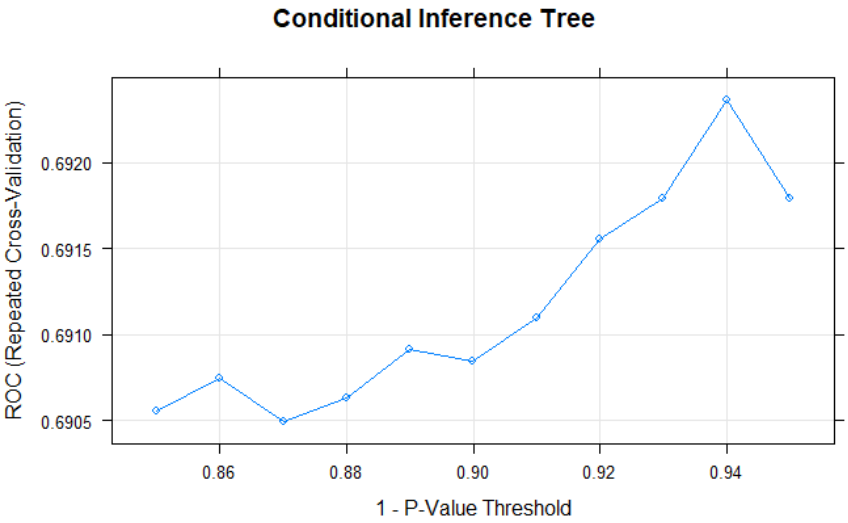

Multivariate Adaptative Regression Splines with VIF<10 predictors
Product Degree
1
2
3
ROC (Repeated Cross-Validation)
0.715
0.710
0.705
0.700
0.695
2
4
6
8
#Terms

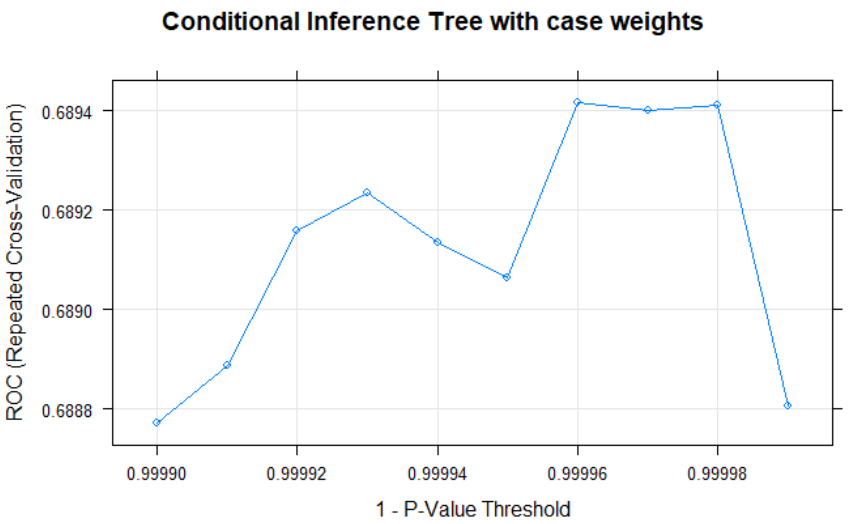

Neural Network with VIF<10 predictors
ROC (Repeated Cross-Validation)
0.71526
0.71524
0.71522
0.71520
0.71518
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Weight Decay

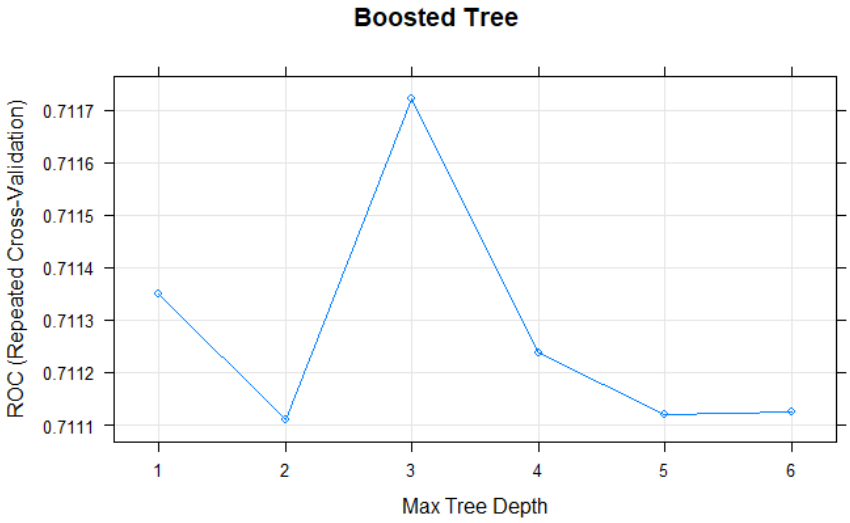

Neural Network with VIF<10 predictors and case weights
ROC (Repeated Cross-Validation)
0.7150
0.7145
0.7140
0.7135
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Weight Decay

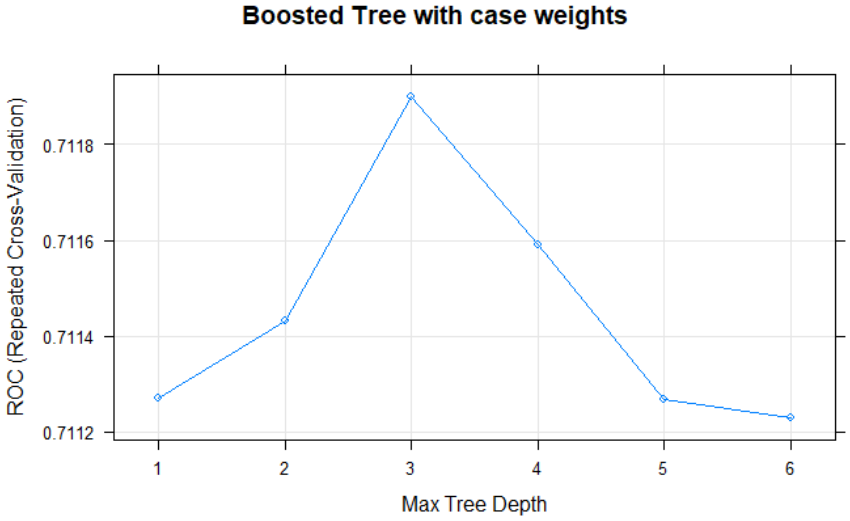

CART
ROC (Repeated Cross-Validation)
0.67
0.66
0.65
0.64
0.63
0.62
0.61
0.005
0.010
0.015
0.020
Complexity Parameter

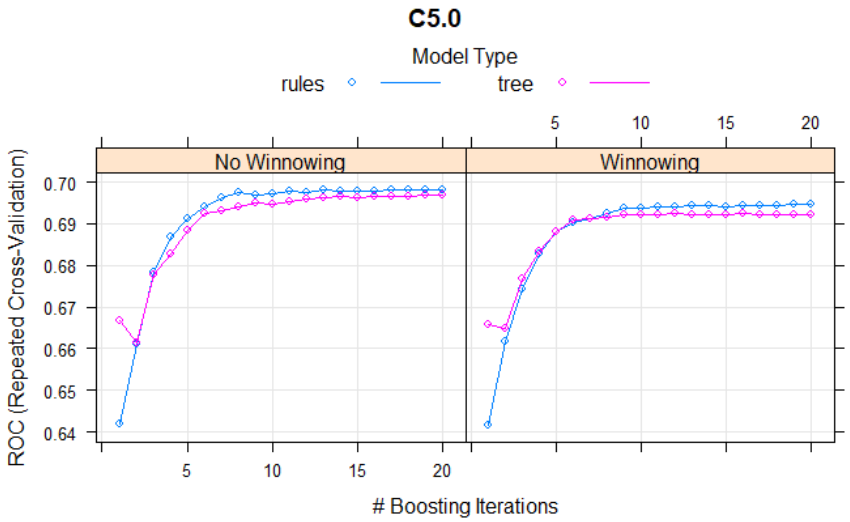

CART with case weights
ROC (Repeated Cross-Validation)
0.66
0.65
0.64
0.63
0.62
0.61
0.005
0.010
0.015
0.020
Complexity Parameter

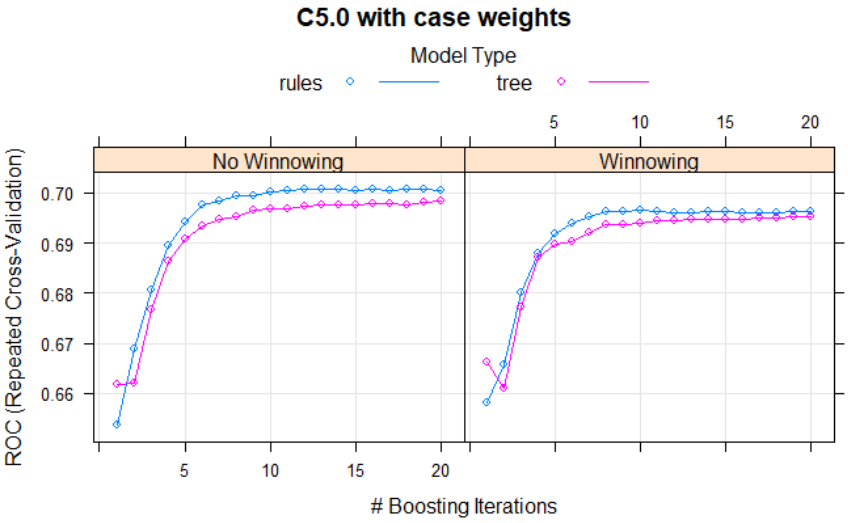

Conditional Inference Tree
ROC (Repeated Cross-Validation)
0.6920
0.6915
0.6910
0.6905
0.86
0.88
0.90
0.92
0.94
1 - P-Value Threshold

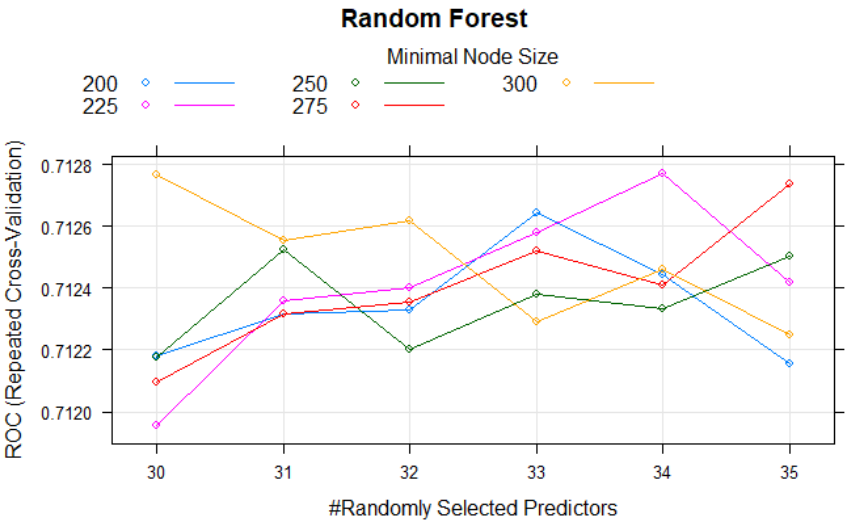

Conditional Inference Tree with case weights
ROC (Repeated Cross-Validation)
0.6894
0.6892
0.6890
0.6888
0.99990
0.99992
0.99994
0.99996
0.99998
1 - P-Value Threshold

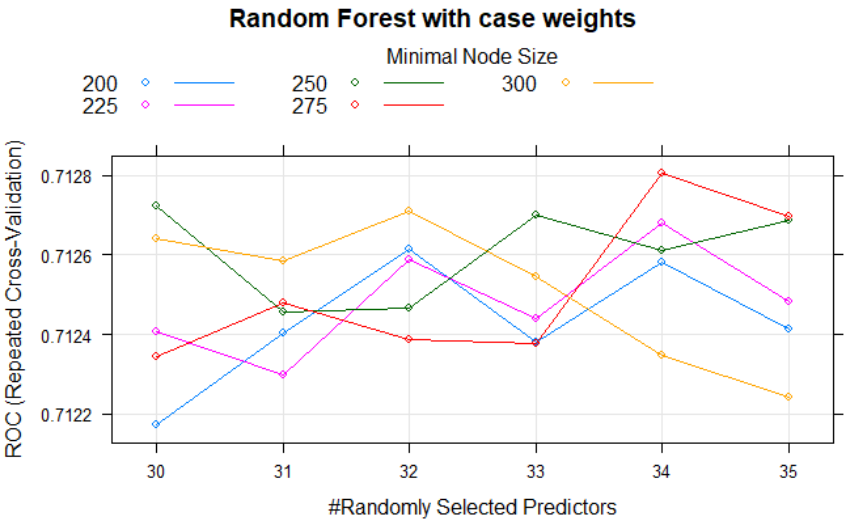

Boosted Tree
ROC (Repeated Cross-Validation)
0.7117
0.7116
0.7115
0.7114
0.7113
0.7112
0.7111
1
2
3
4
5
6
Max Tree Depth

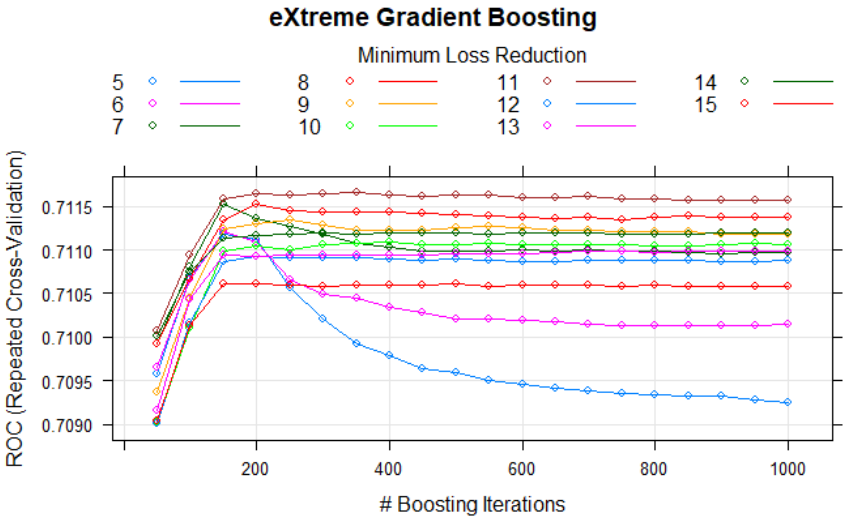

Boosted Tree with case weights
ROC (Repeated Cross-Validation)
0.7118
0.7116
0.7114
0.7112
1
2
3
4
5
6
Max Tree Depth

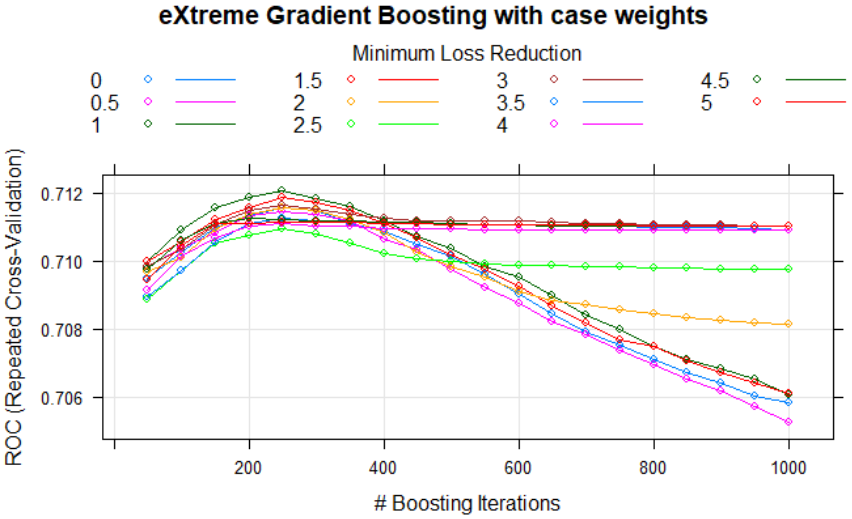

C5.0
Model Type
rules
tree
No Winnowing
Winnowing
ROC (Repeated Cross-Validation)
0.70
0.69
0.68
0.67
0.66
0.65
0.64
5
10
15
20
Boosting Iterations

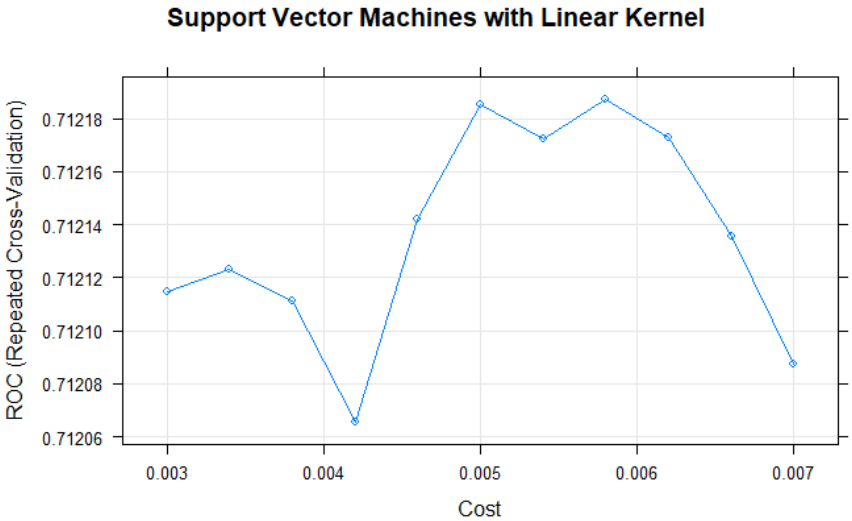

C5.0 with case weights
Model Type
rules
tree
No Winnowing
Winnowing
ROC (Repeated Cross-Validation)
Boosting Iterations


Random Forest
Minimal Node Size
200
225
250
275
300
ROC (Repeated Cross-Validation)
#Randomly Selected Predictors

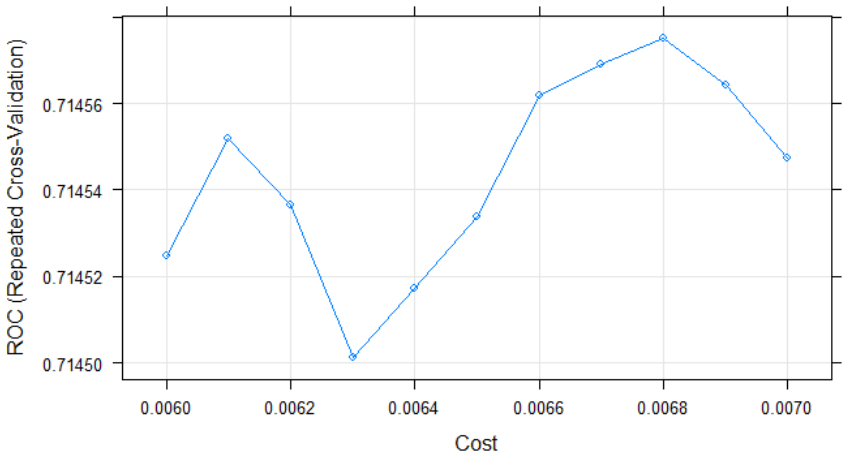

Random Forest with case weights
Minimal Node Size
200
225
250
275
300
ROC (Repeated Cross-Validation)
#Randomly Selected Predictors


eXtreme Gradient Boosting
Minimum Loss Reduction
5
6
7
8
9
10
11
12
13
14
15
ROC (Repeated Cross-Validation)
0.7115
0.7110
0.7105
0.7100
0.7095
0.7090
200
400
600
800
1000
Boosting Iterations

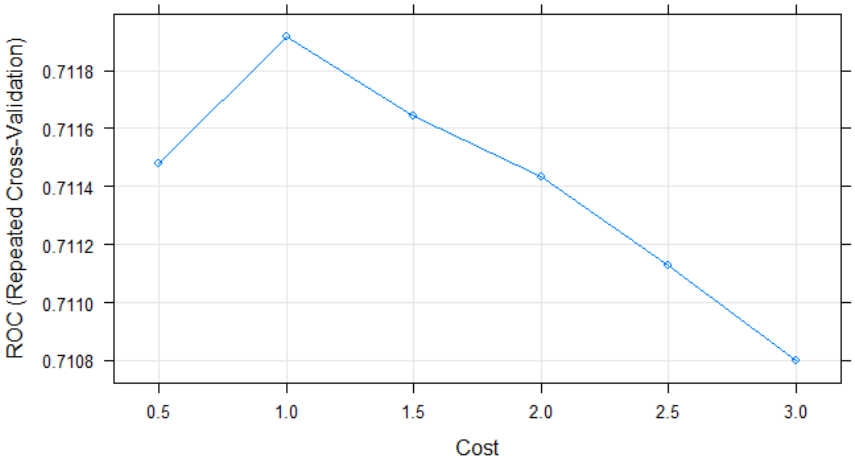

eXtreme Gradient Boosting with case weights
Minimum Loss Reduction
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
ROC (Repeated Cross-Validation)
0.712
0.710
0.708
0.706
200
400
600
800
1000
Boosting Iterations


Support Vector Machines with Linear Kernel
ROC (Repeated Cross-Validation)
0.71218
0.71216
0.71214
0.71212
0.71210
0.71208
0.71206
0.003
0.004
0.005
0.006
0.007
Cost

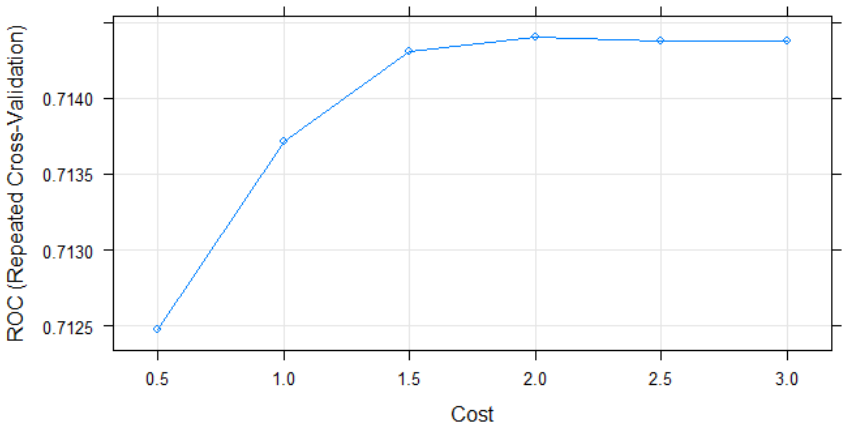
Support Vector Machines with Linear Kernel with VIF<10 predictors

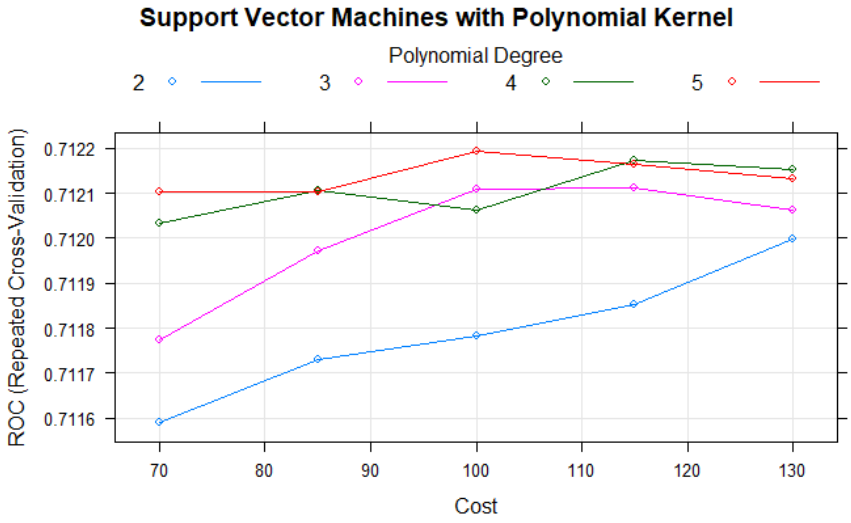


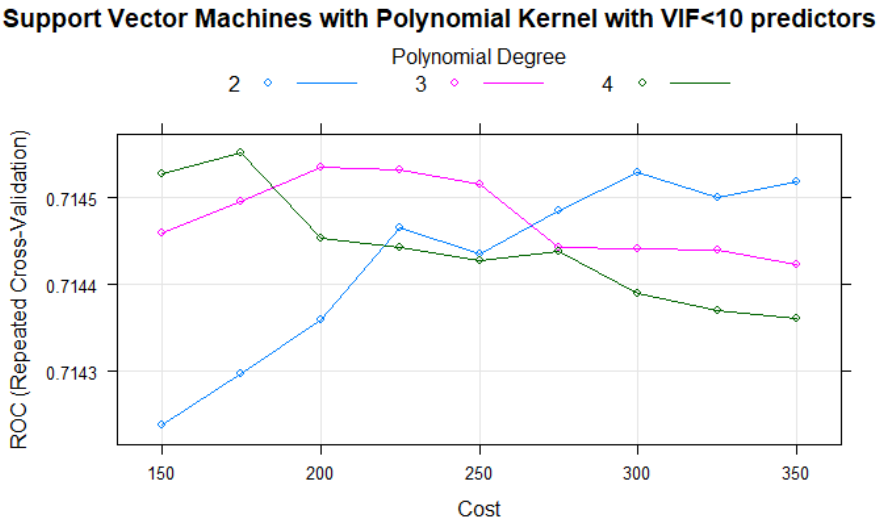
Support Vector Machines with Radial Basis Function Kernel



Support Vector Machines with Radial Basis Function Kernel with VIF<10 predictors

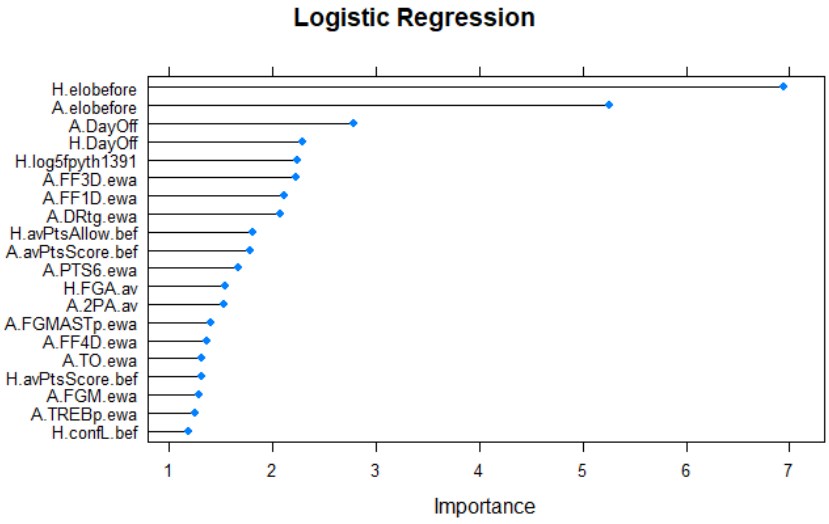


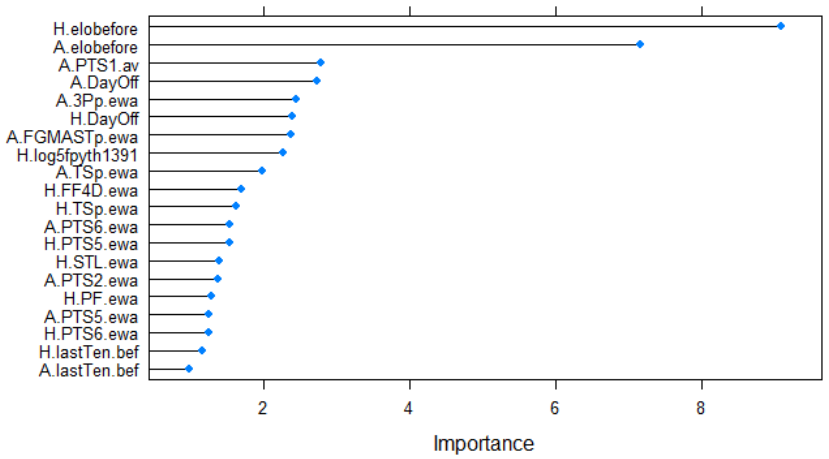

Support Vector Machines with Polynomial Kernel
Polynomial Degree
2
3
4
5
ROC (Repeated Cross-Validation)
0.7122
0.7121
0.7120
0.7119
0.7118
0.7117
0.7116
70
80
90
100
110
120
130
Cost

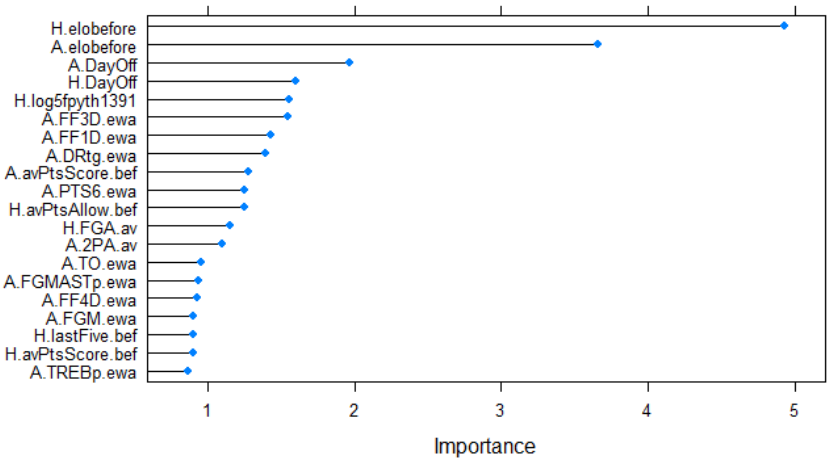

Support Vector Machines with Polynomial Kernel with VIF<10 predictors
Polynomial Degree
2
3
4
ROC (Repeated Cross-Validation)
0.7145
0.7144
0.7143
150
200
250
300
350
Cost

Appendix B

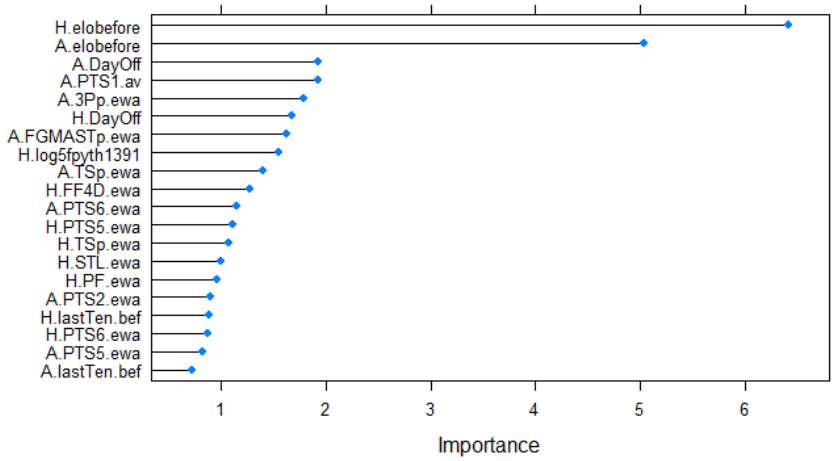
Variable Importance Plot



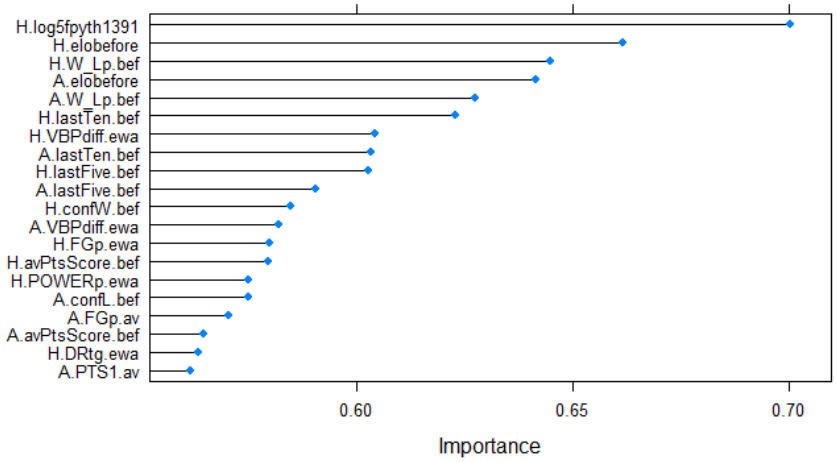




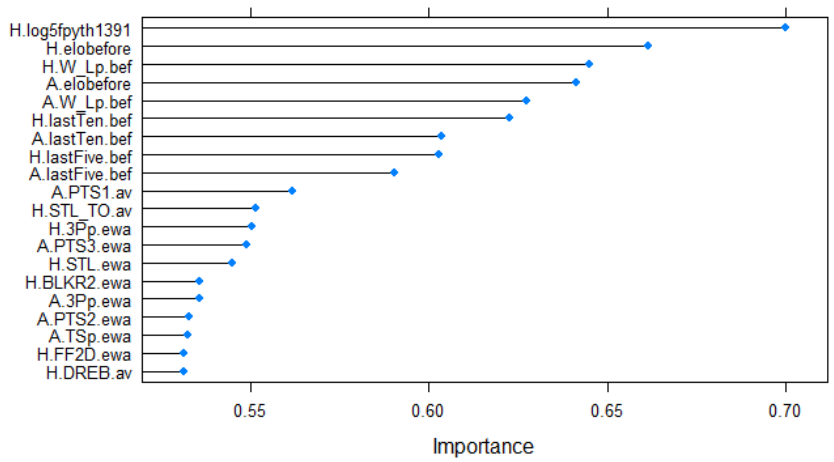
Logistic Regression with VIF<10 predictors and case weights



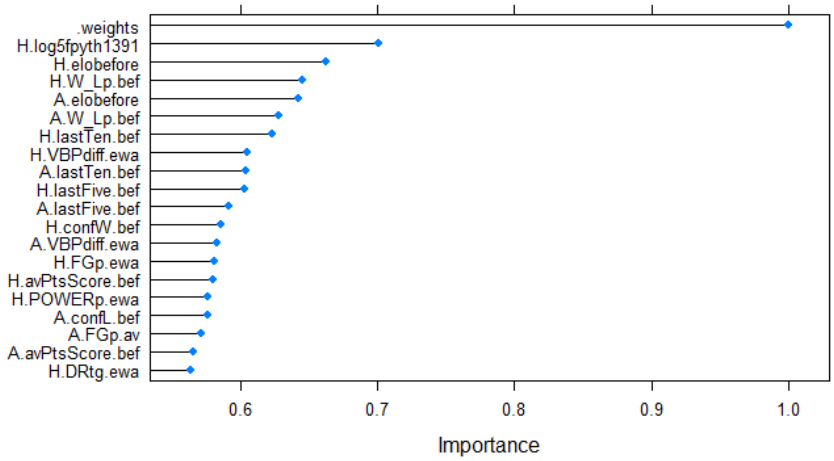
Bayesian Logistic Regression



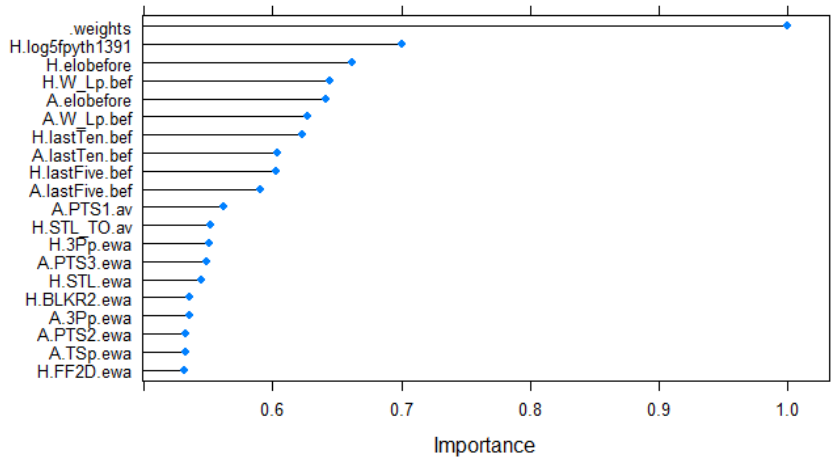
Bayesian Logistic Regression with VIF<10 predictors



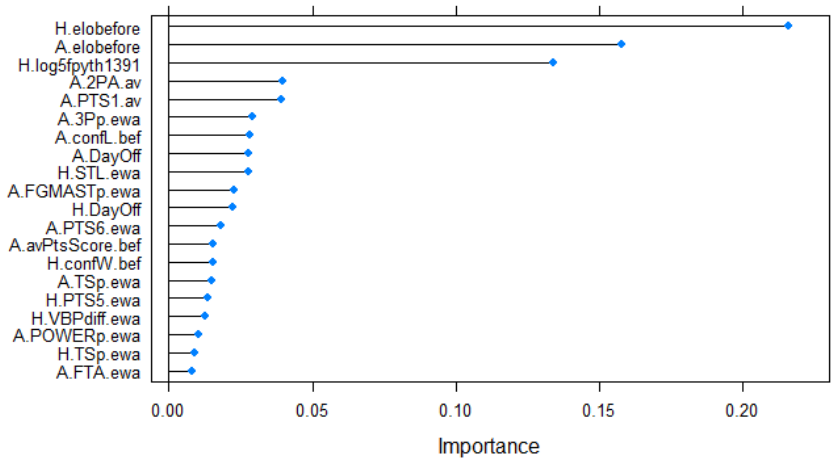
Bayesian Logistic Regression with case weights



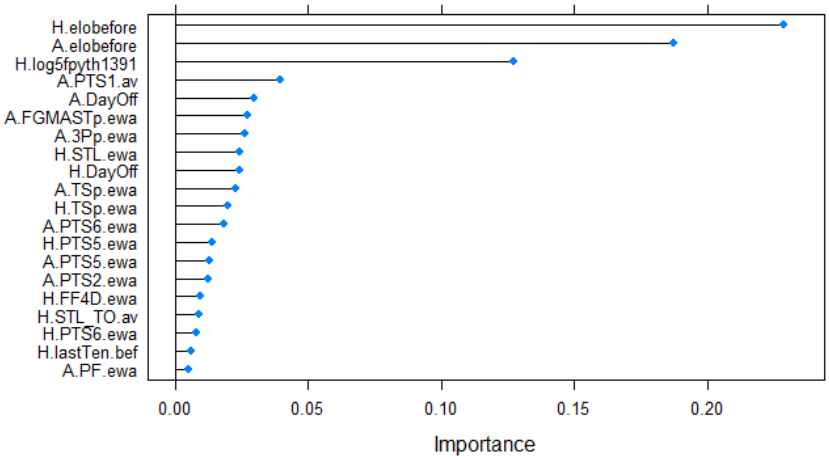
Bayesian Logistic Regression with VIF<10 predictors and case weights



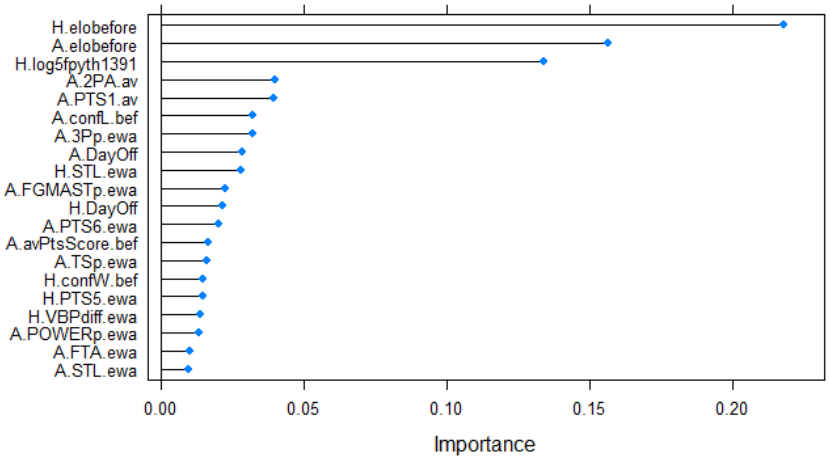
Boosted Logistic Regression



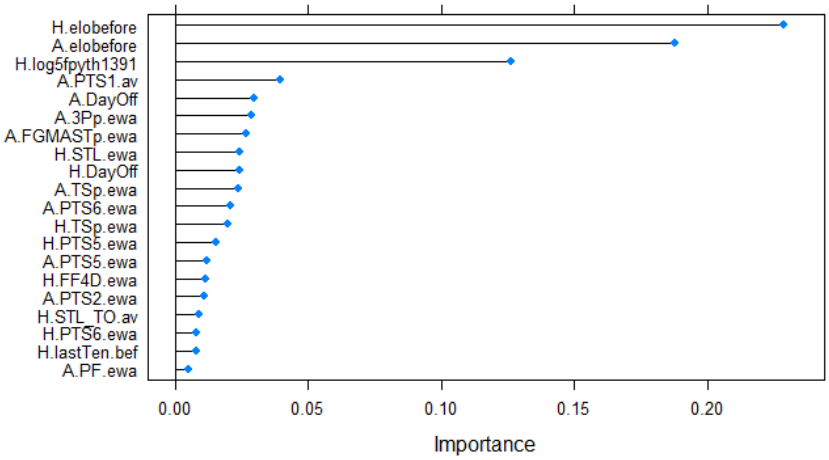
Boosted Logistic Regression with VIF<10 predictors



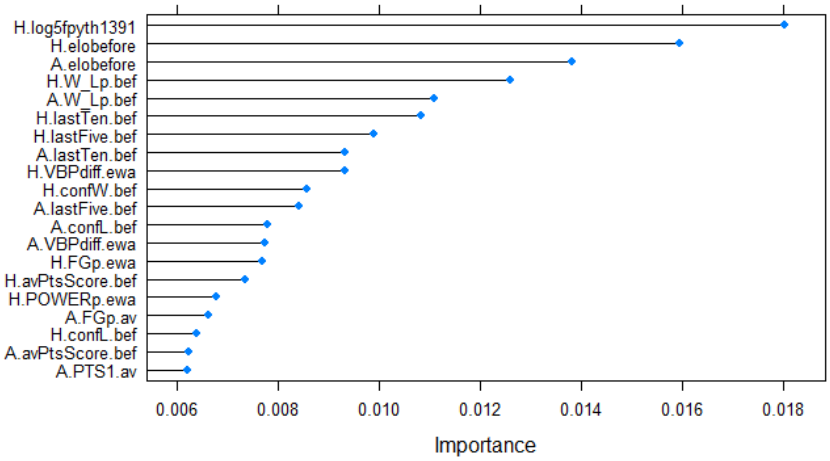
Boosted Logistic Regression with case weights



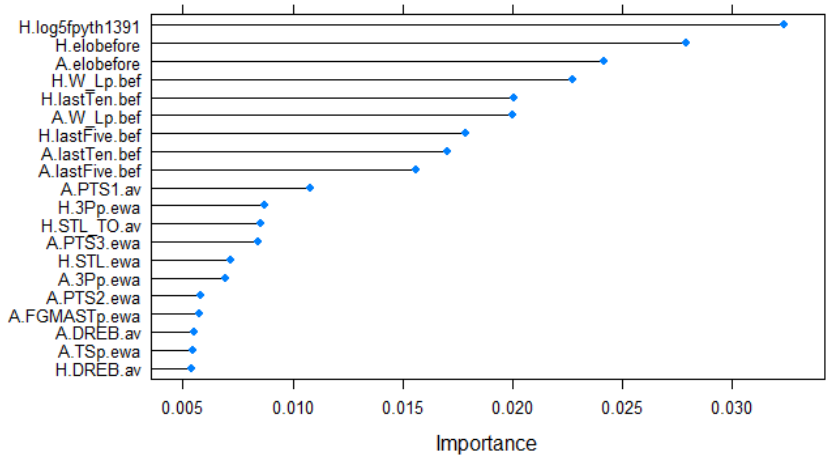
Boosted Logistic Regression with VIF<10 predictors and case weights



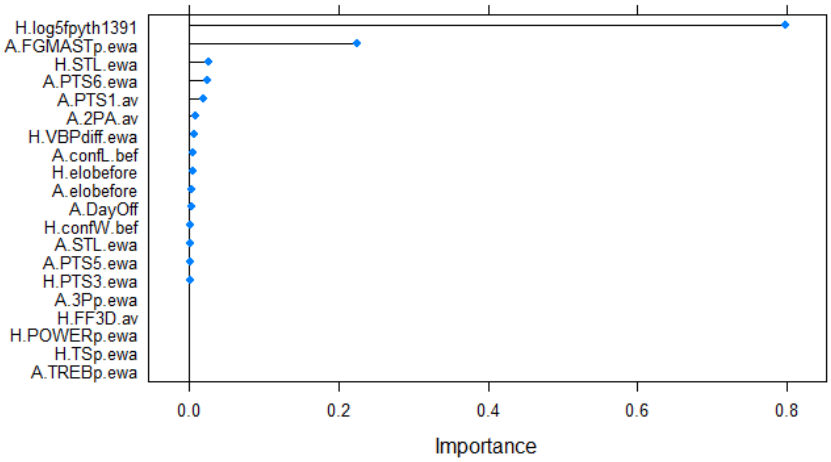
Partial Least Squares Regression



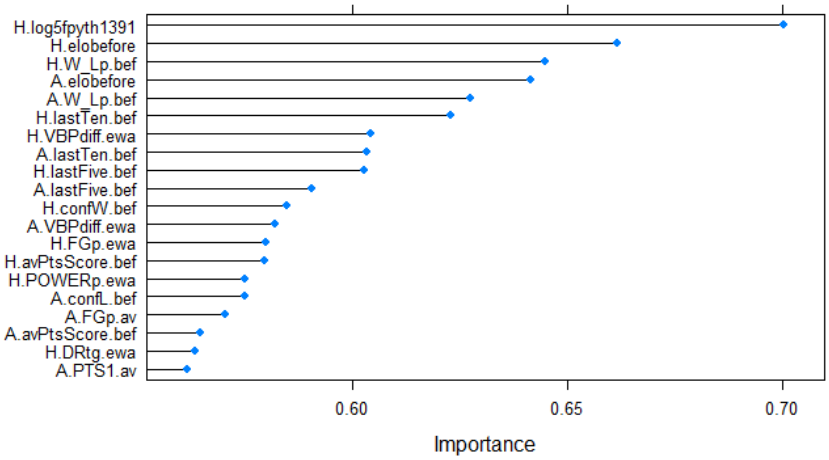
Partial Least Squares Regression with VIF<10 predictors



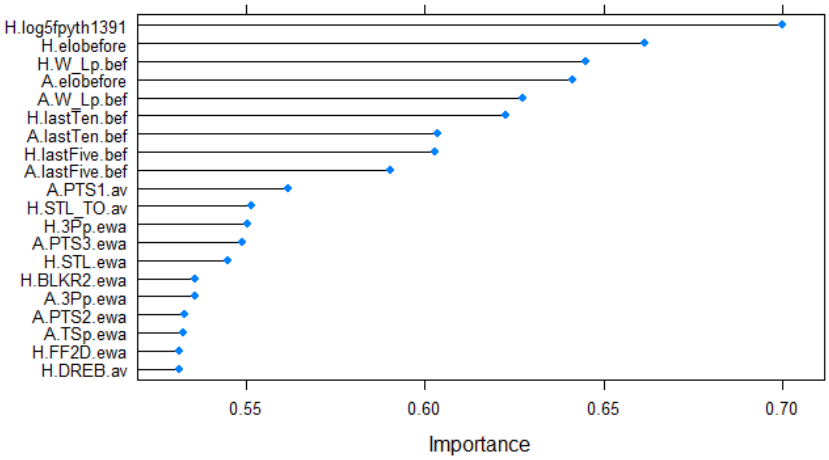
Glmnet



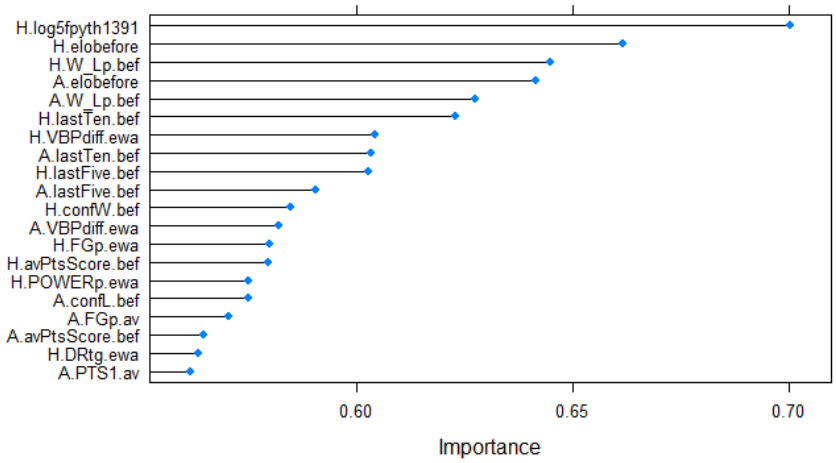
Linear Discriminant Analysis



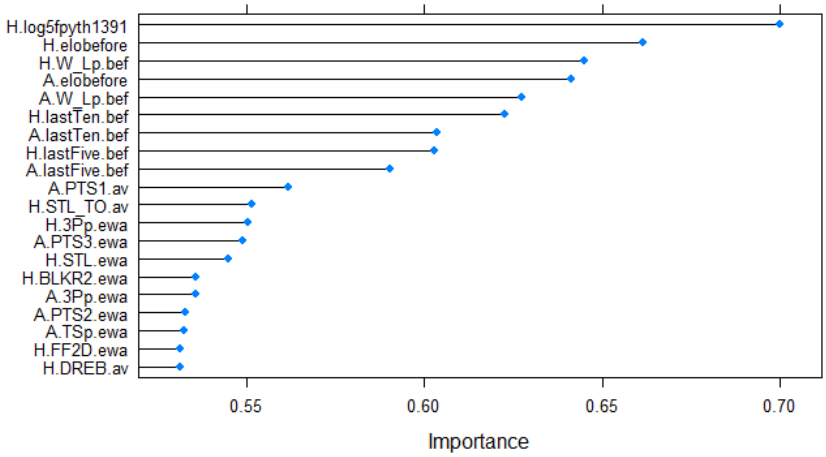
Linear Discriminant Analysis with VIF<10 predictors



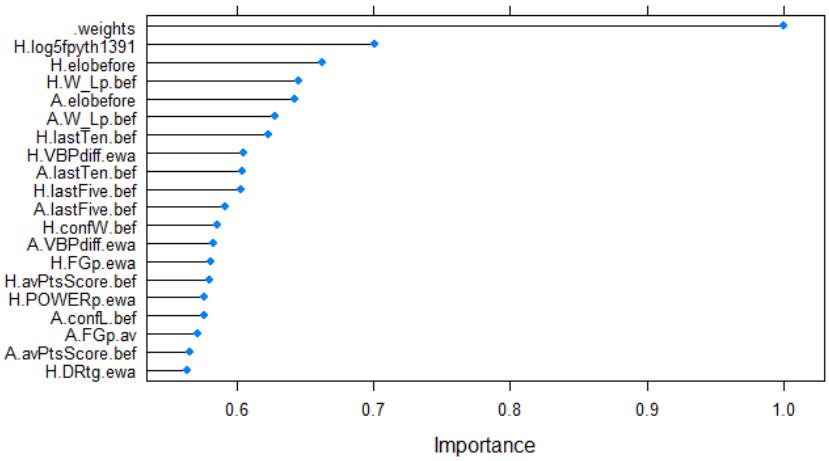
Penalized Discriminant Analysis



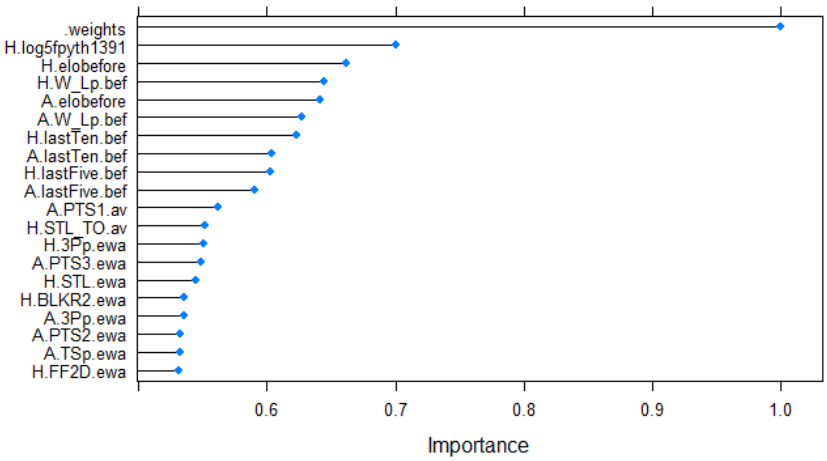
Penalized Discriminant Analysis with VIF<10 predictors



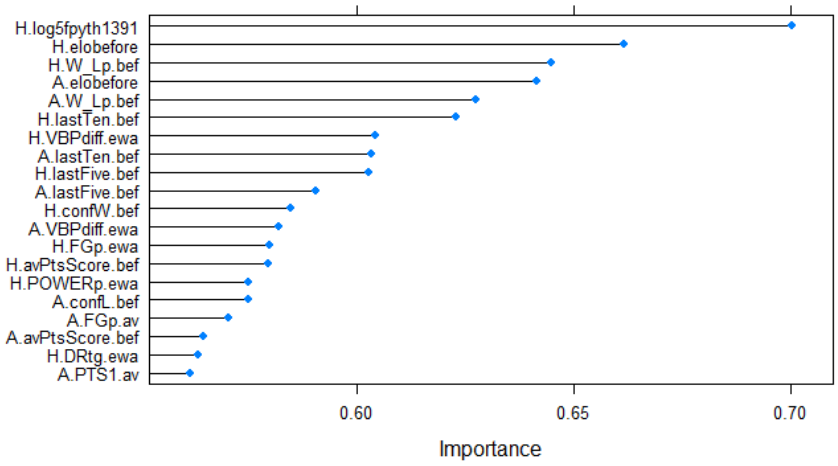
Penalized Discriminant Analysis with case weights



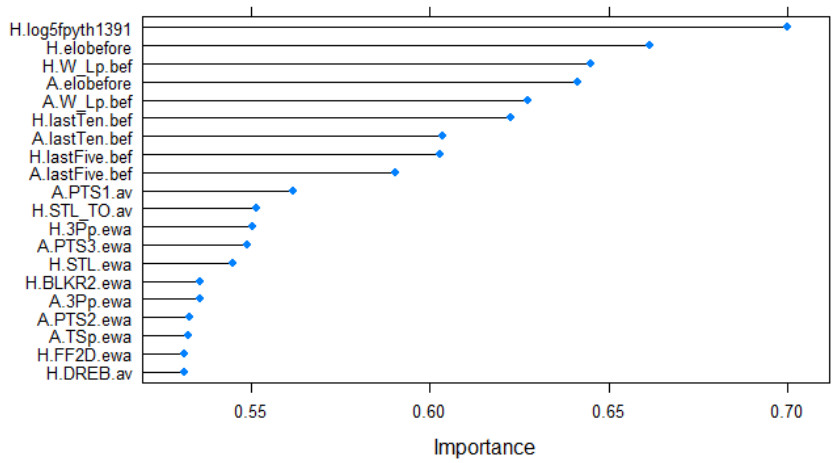
Penalized Discriminant Analysis with VIF<10 predictors and case weights



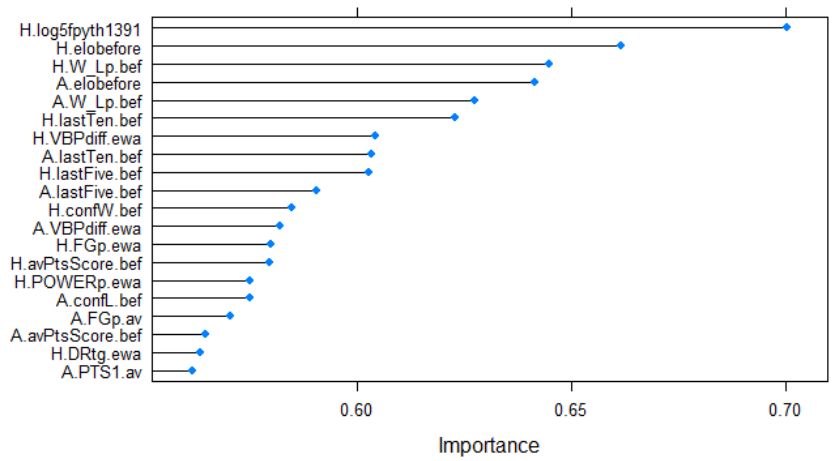
Regularized Discriminant Analysis



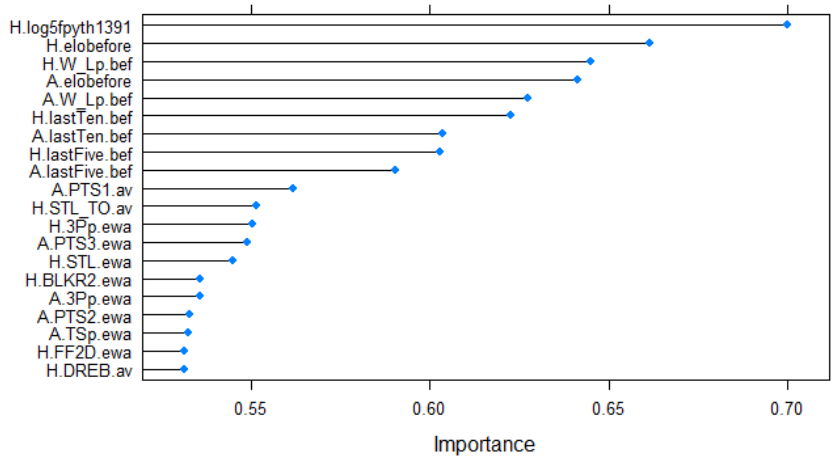
Regularized Discriminant Analysis with VIF<10 predictors



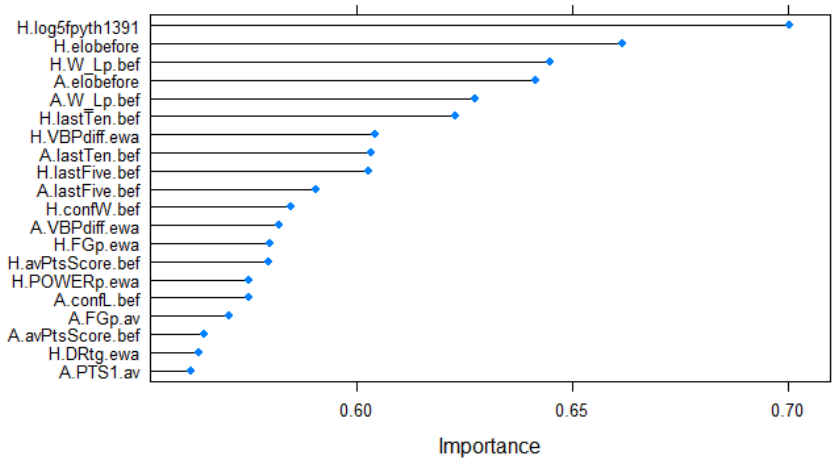
K-Nearest Neighbours



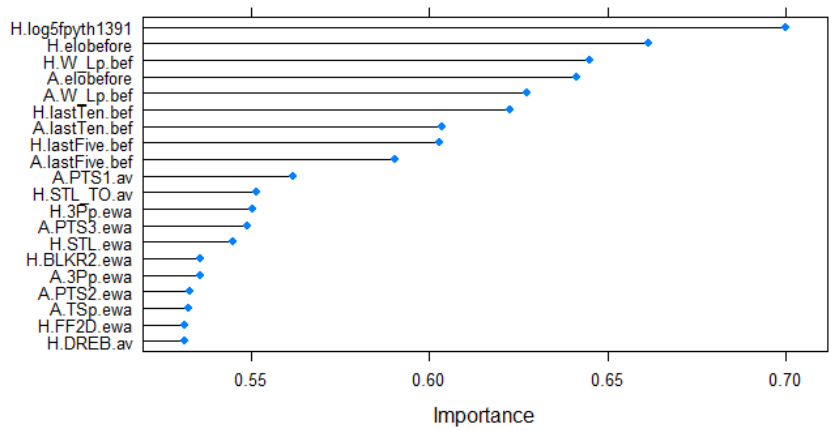
K-Nearest Neighbours with VIF<10 predictors



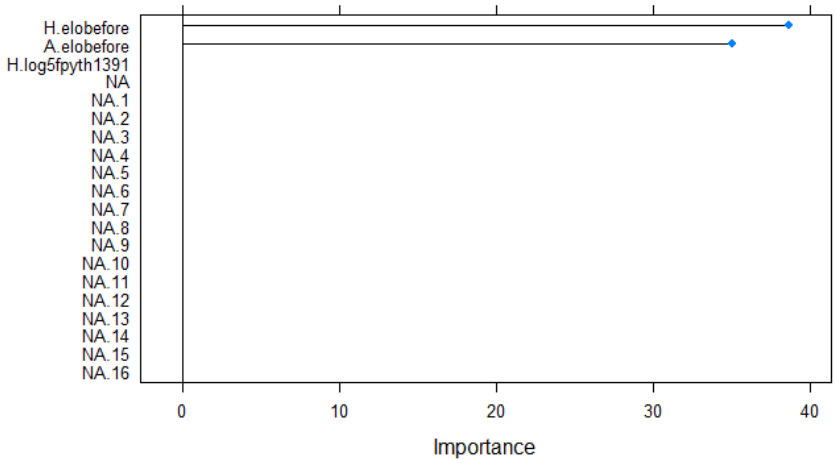
K-Nearest Neighbours with Principal Components Analysis



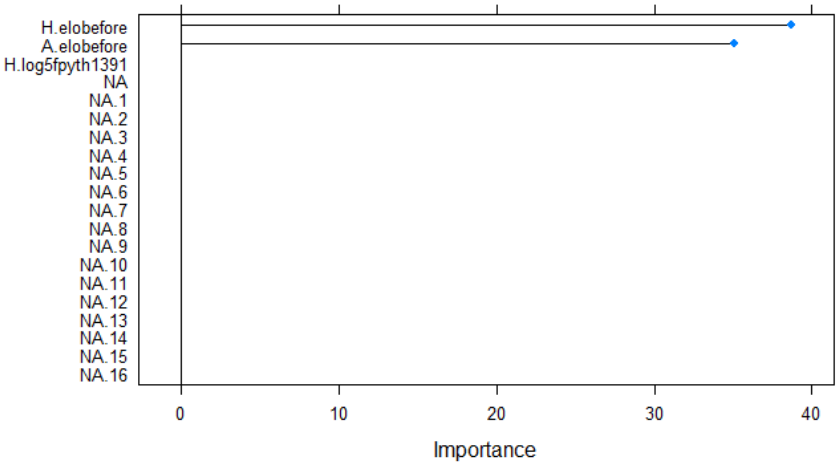
K-Nearest Neighbours with VIF<10 predictors and Principal Components Analysis



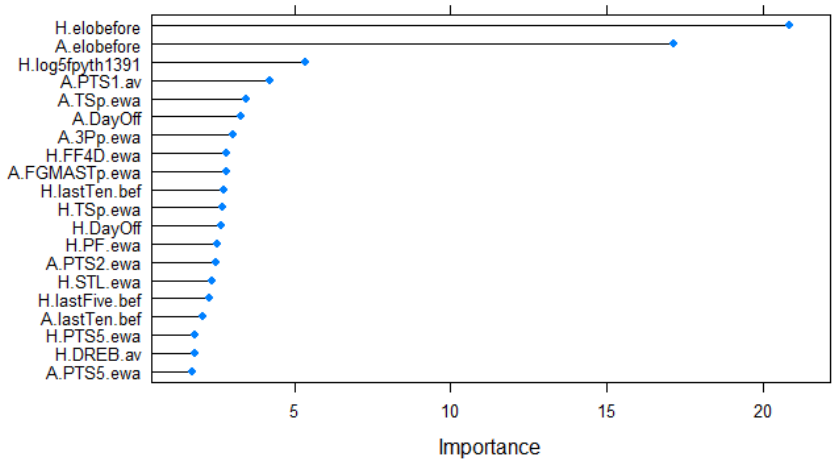
Multivariate Adaptative Regression Splines



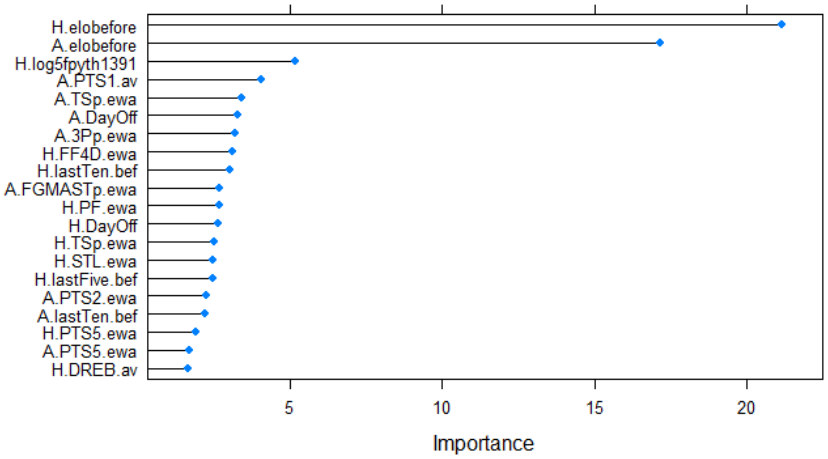
Multivariate Adaptative Regression Splines with VIF<10 predictors



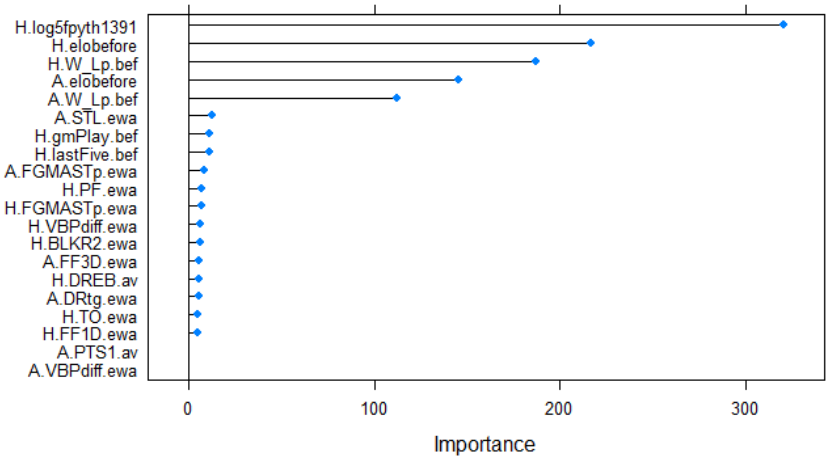
Neural Network with VIF<10 predictors



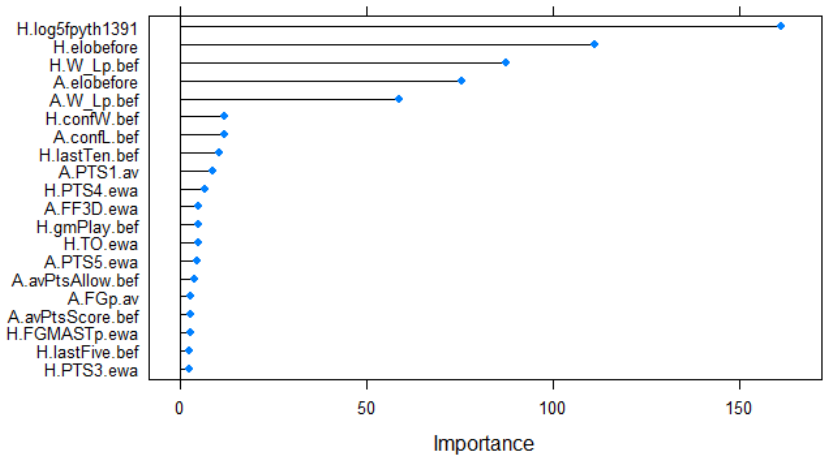
Neural Network with VIF<10 predictors and case weights



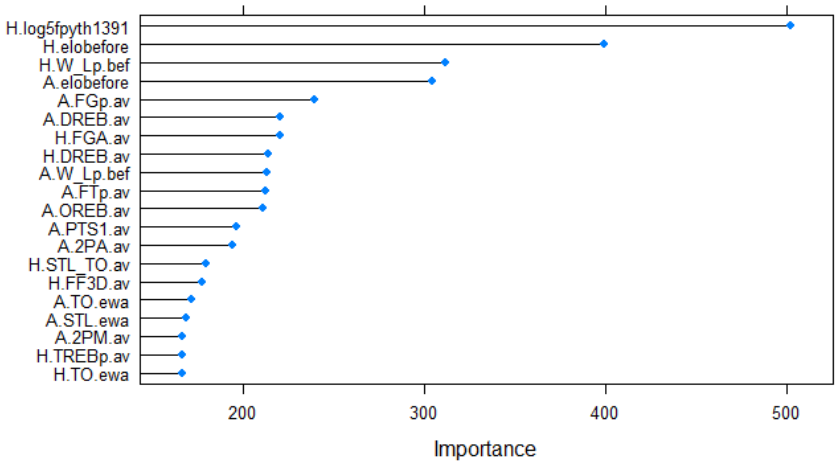
CART



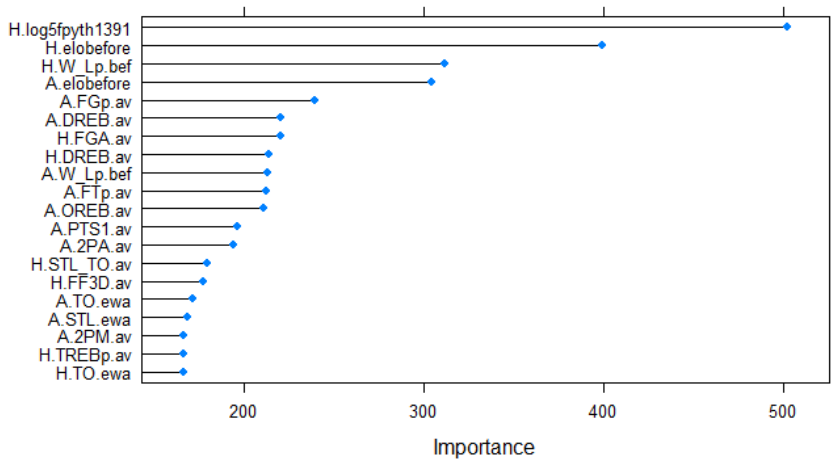
CART with case weights



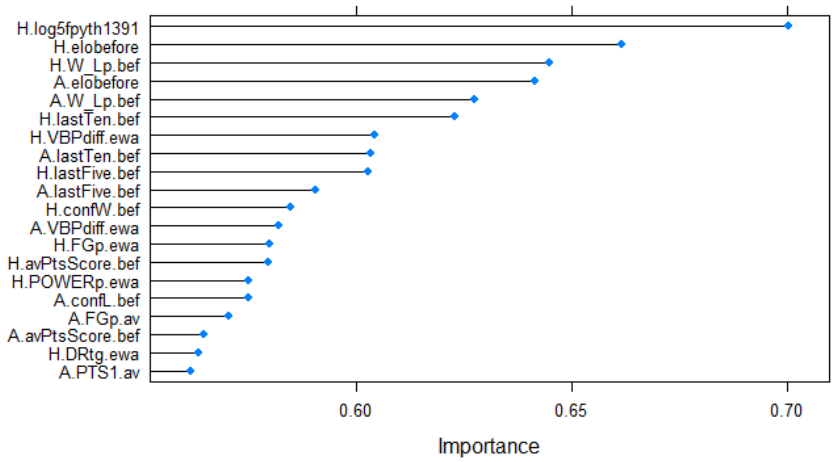
Bagged CART

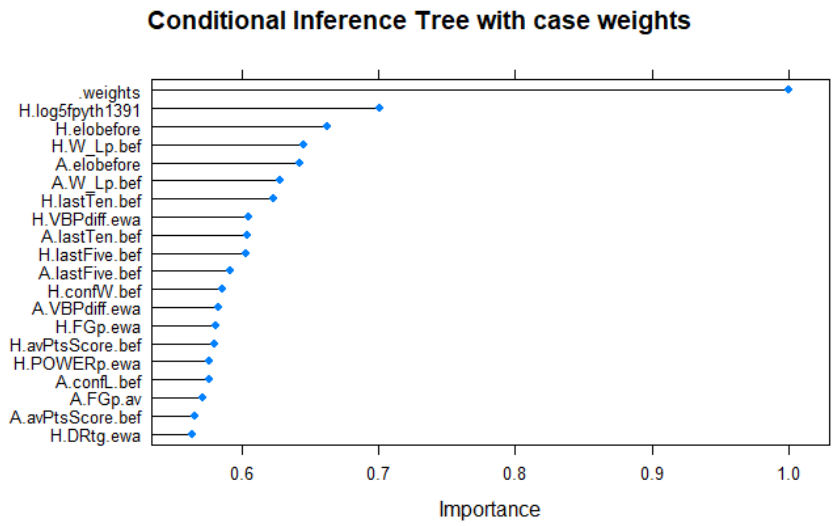


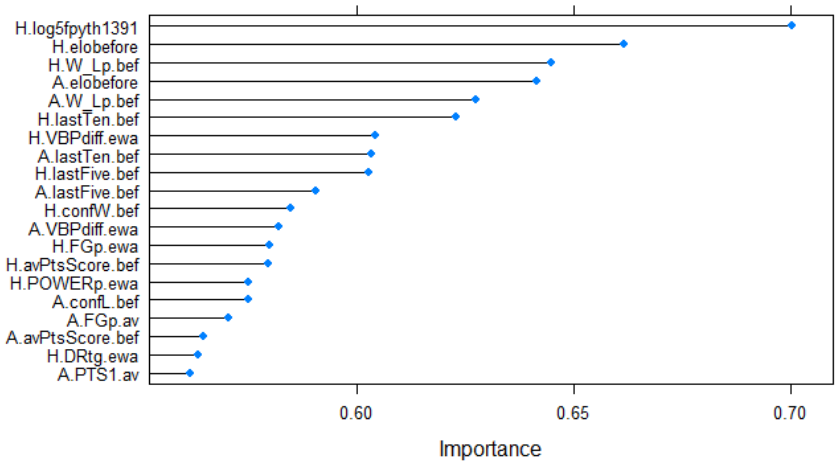
Bagged CART with case weights

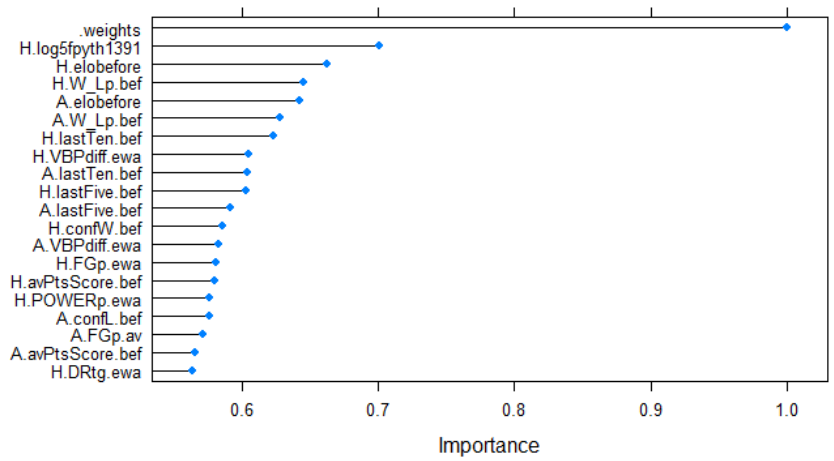


Conditional Inference Tree

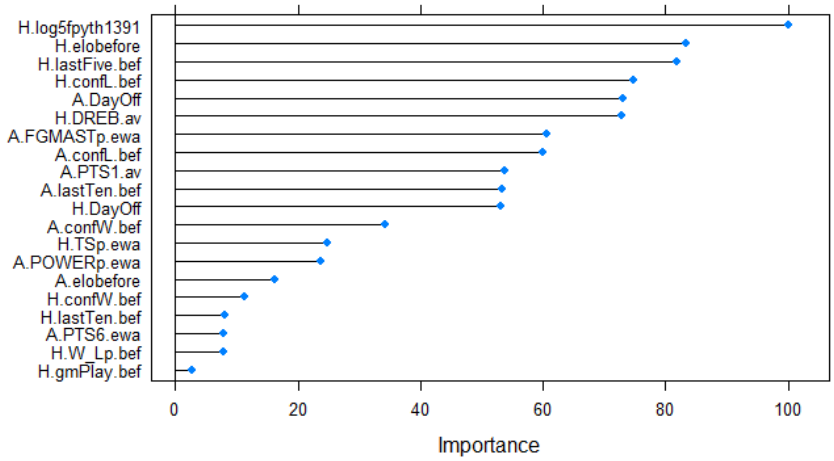



Conditional Inference Tree with case weights
.weights
H.log5fpyth1391
H.elobefore
H.W_Lp.bef
A.elobefore
A.W_Lp.bef
H.lastTen.bef
H.VBPdiff.ewa
A.lastTen.bef
H.lastFive.bef
A.lastFive.bef
H.confW.bef
A.VBPdiff.ewa
H.FGp.ewa
H.avPtsScore.bef
H.POWERp.ewa
A.confL.bef
A.FGp.av
A.avPtsScore.bef
H.DRtg.ewa
0.6
0.7
0.8
0.9
1.0
Importance

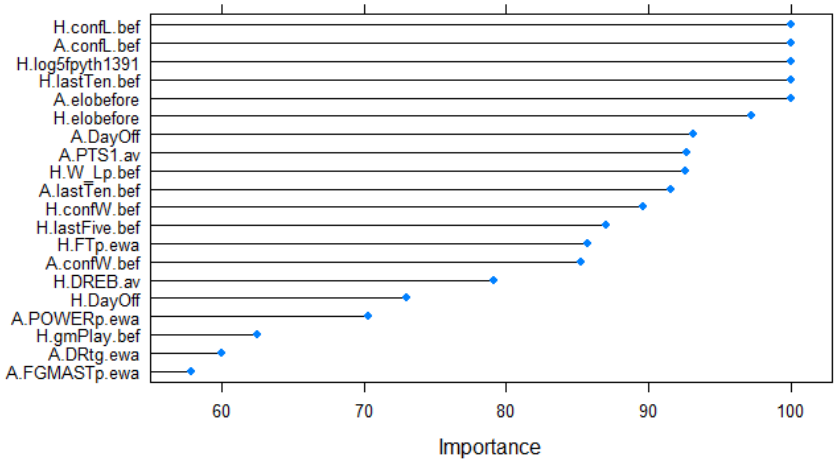

Boosted Tree
H.log5fpyth1391
H.elobefore
H.W_Lp.bef
A.elobefore
A.W_Lp.bef
H.lastTen.bef
H.VBPdiff.ewa
A.lastTen.bef
H.lastFive.bef
A.lastFive.bef
H.confW.bef
A.VBPdiff.ewa
H.FGp.ewa
H.avPtsScore.bef
H.POWERp.ewa
A.confL.bef
A.FGp.av
A.avPtsScore.bef
H.DRtg.ewa
A.PTS1.av
0.60
0.65
0.70
Importance


Boosted Tree with case weights
.weights
H.log5fpyth1391
H.elobefore
H.W_Lp.bef
A.elobefore
A.W_Lp.bef
H.lastTen.bef
H.VBPdiff.ewa
A.lastTen.bef
H.lastFive.bef
A.lastFive.bef
H.confW.bef
A.VBPdiff.ewa
H.FGp.ewa
H.avPtsScore.bef
H.POWERp.ewa
A.confL.bef
A.FGp.av
A.avPtsScore.bef
H.DRtg.ewa
0.6
0.7
0.8
0.9
1.0
Importance

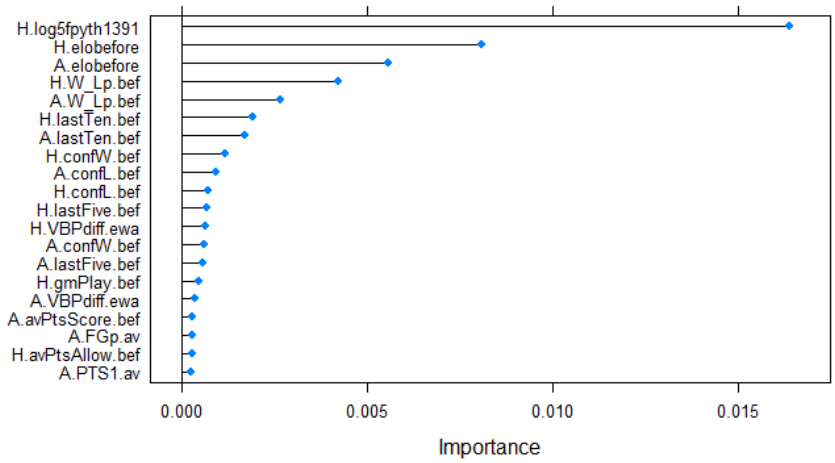
C5.0

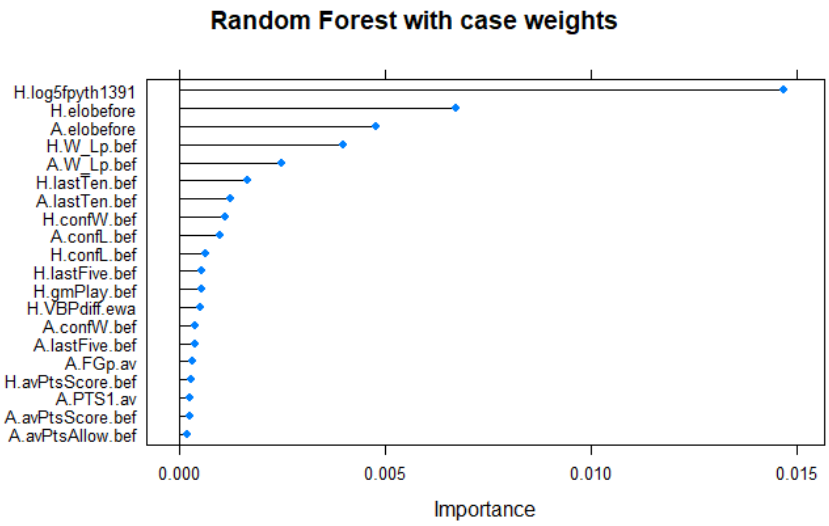


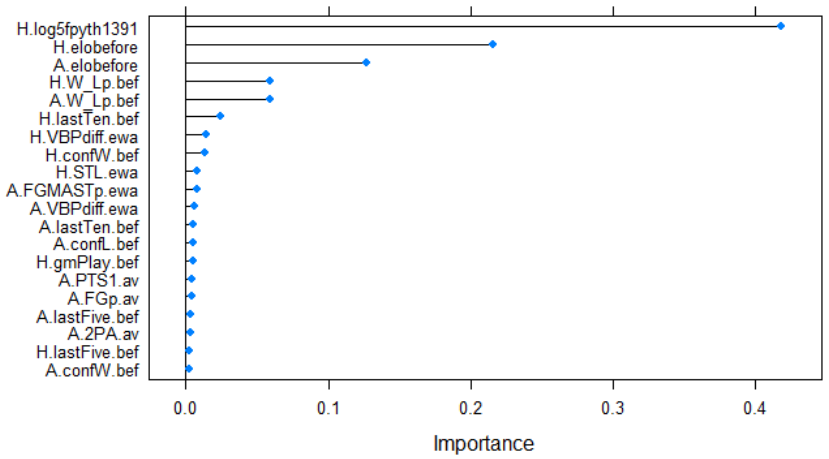
C5.0 with case weights

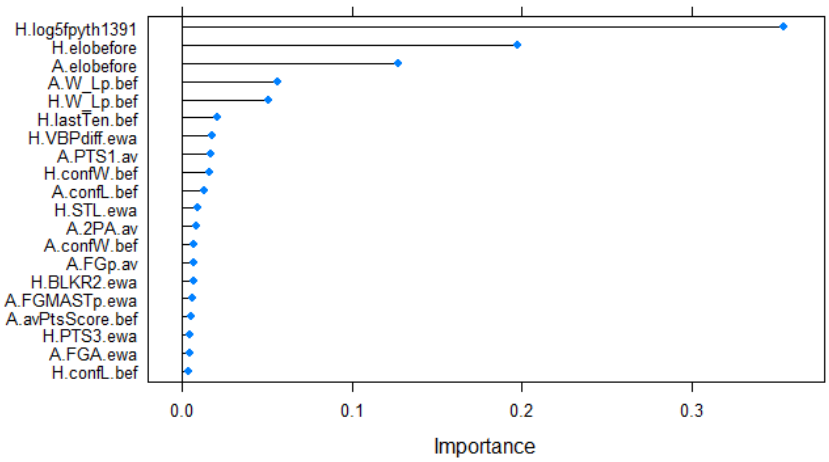


Random Forest

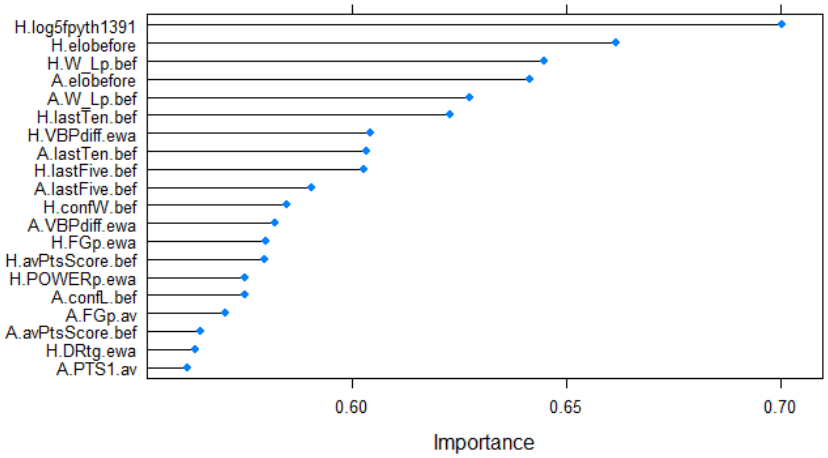



Random Forest with case weights
H.log5fpyth1391
H.elobefore
A.elobefore
H.W_Lp.bef
A.W_Lp.bef
H.lastTen.bef
A.lastTen.bef
H.confW.bef
A.confL.bef
H.confL.bef
H.lastFive.bef
H.gmPlay.bef
H.VBPdiff.ewa
A.confW.bef
A.lastFive.bef
A.FGp.av
H.avPtsScore.bef
A.PTS1.av
A.avPtsScore.bef
A.avPtsAllow.bef
0.000
0.005
0.010
0.015
Importance

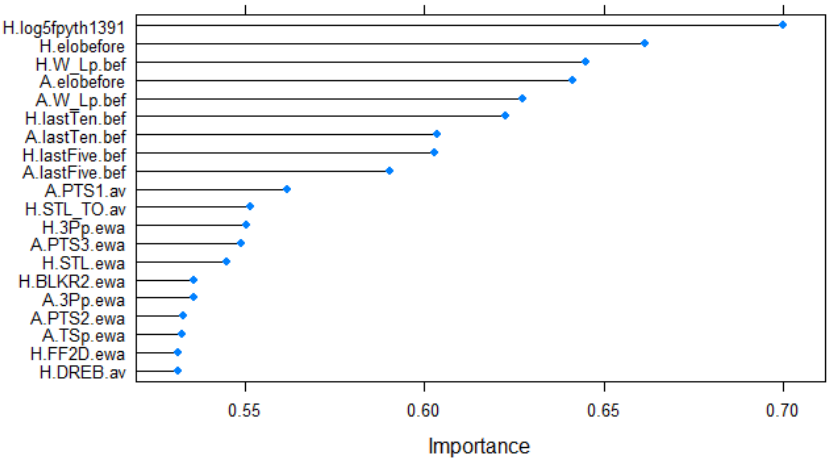

eXtreme Gradient Boosting
H.log5fpyth1391
H.elobefore
A.elobefore
H.W_Lp.bef
A.W_Lp.bef
H.lastTen.bef
H.VBPdiff.ewa
H.confW.bef
H.STL.ewa
A.FGMASTp.ewa
A.VBPdiff.ewa
A.lastTen.bef
A.confL.bef
H.gmPlay.bef
A.PTS1.av
A.FGp.av
A.lastFive.bef
A.2PA.av
H.lastFive.bef
A.confW.bef
0.0
0.1
0.2
0.3
0.4
Importance


eXtreme Gradient Boosting with case weights
H.log5fpyth1391
H.elobefore
A.elobefore
A.W_Lp.bef
H.W_Lp.bef
H.lastTen.bef
H.VBPdiff.ewa
A.PTS1.av
H.confW.bef
A.confL.bef
H.STL.ewa
A.2PA.av
A.confW.bef
A.FGp.av
H.BLKR2.ewa
A.FGMASTp.ewa
A.avPtsScore.bef
H.PTS3.ewa
A.FGA.ewa
H.confL.bef
0.0
0.1
0.2
0.3
Importance

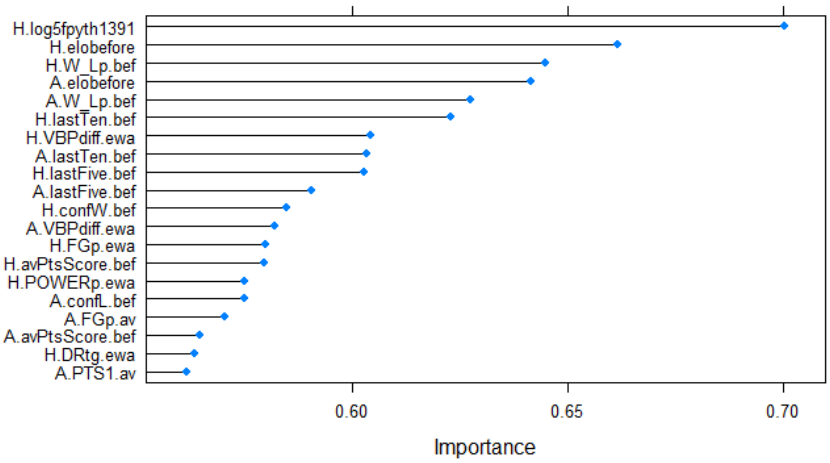
Support Vector Machines with Linear Kernel



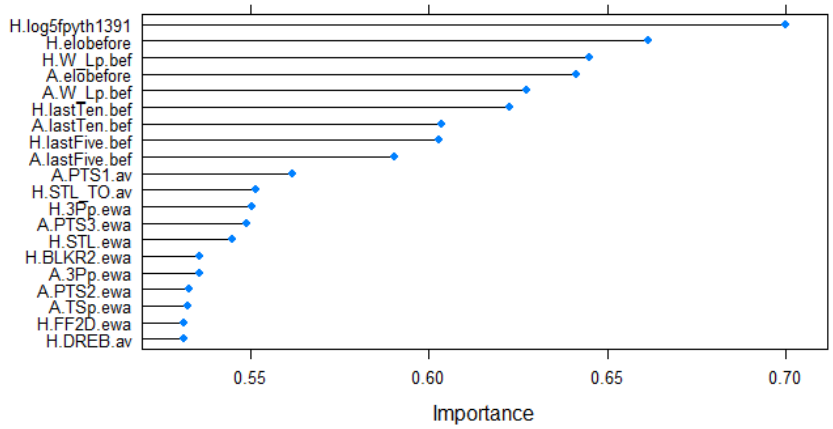
Support Vector Machines with Linear Kernel with VIF<10 predictors



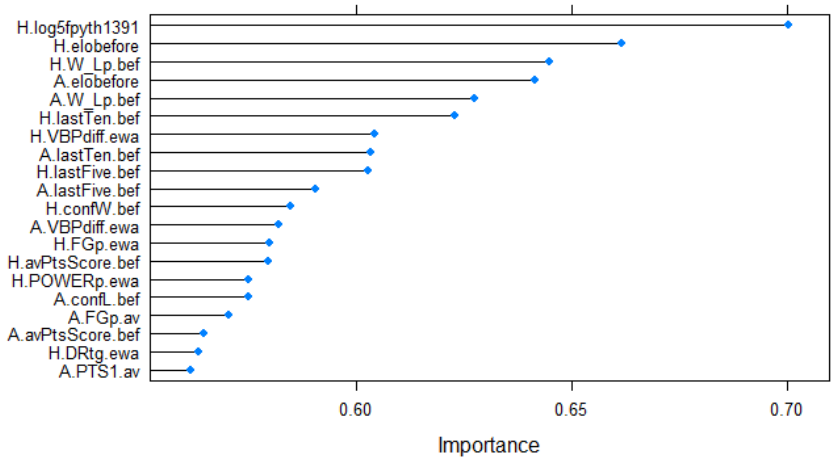
Support Vector Machines with Radial Basis Function Kernel



Support Vector Machines with Radial Basis Function Kernel with VIF<10 predictors



Support Vector Machines with Polynomial Kernel



Support Vector Machines with Polynomial Kernel with VIF<10 predictors