Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis: Data Science and Marketing Analytics

Analyzing the predictive power of advanced and traditional statistics in football

Michaël Matthews
Student ID: 407955
Supervisor: Nierop, J.E.M. van
Second Assessor: Fok, D.
Date: 25-04-2022

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Table of Contents

Abstract

Football is perhaps the world's most popular sport and there is no shortage of match statistics available from every match that was played. Recently advanced statistics are becoming increasingly more popular. Advanced statistics try to show match statistics on a deeper level than a regular fan would notice or that could be seen in a regular box score after a match. Which is why this research aims to determine whether advanced statistics can be used to build a better performing prediction model than traditional statistics for football match results. This is done by predicting future matches using statistics, both traditional and advanced, from previous matches in machine learning models. And afterwards comparing the results obtained from the two types of statistics. Previous research has attempted to predict football match results utilizing different datasets and variables. Three different datasets were created for every kind of statistics containing the results of the past five seasons of the English, Spanish, German, French and Italian Leagues to answer the research question. These datasets were then used in three different machine learning models: Multinomial Logistics Regression, Random Forest and Artificial Neural Network, to determine which type of statistics returned the better-perfoming models. The models' performances were measured using three metrics: accuracy, F-measure and Ranked Probability Score. In addition, these models helped determine the most important individual statistic from these datasets. multinomial logistic regression returned the highest overall accuracy with 54.78%, the highest F-measure was reached with traditional statistics at 0.429 with a random forest model. The best performing model according to RPS score was 0.2021, reached with advanced statistics using a multinomial logistic regression model. These datasets also included highly influential variables in every model, such as SPI for the advanced statistics. From the research results, it can be concluded that using advanced statistics results in a better-performing model for predicting football match results.

Keywords: Football, Advanced Statistics, Traditional Statistics, Machine Learning, Classification

1 Introduction

Football or soccer, as it is also known, can be considered one of the most popular team sports in the world. For example, more than half a billion watched the 2018 World Cup final (Richter, 2021), ironically the least viewed of the last 3 World Cup finals (Richter, 2021). However, football is difficult to predict despite how much fans or bookmakers want to, even when a much stronger team plays against a weaker opponent. This can be explained due to outcomes in football being generally very low, for example, 1-0 or 2-1, compared to, for instance, basketball, where there is a higher score count (Robberechts et al., 2019). The low-scoring nature of the sport causes the game to be influenced much more by chance (Robberechts et al., 2019). This implies that match results frequently do not reflect the reality of how the game transpired on the field. This is where advanced statistics might be able to step in and give a better indication of the performance displayed.

But what are advanced statistics when it comes to sports matches? Advanced statistics are the new wave of the sports world. In football, the traditional statistics are goals, assists and perhaps possession percentage. But do all these count the same without including other factors around them? That is what advanced statistics try to determine. Advanced statistics try to go deeper than the traditional statistics mentioned before, and they try to detect subtle intricacies of matches that traditional statistics might not consider or even notice (Giasemidis, 2020). This is all possible because of an increase in the data taken out of football (and sports in general) nowadays, which has made the use of advanced statistics in football grow exponentially (Fernandez et al., 2021). Advanced statistics can show the elusive moments in the game that traditional statistics may not be able to capture, or it can determine the brilliance of a player who might not necessarily score goals every game or provide many assists (Giasemidis, 2020). Advanced statistics have proven successful, with teams winning their respective leagues with teams built on analytics (Schoenfield, 2019 & Pratley, 2020).

There are quite a few advanced statistics in football, but some are more common than others. One of the most common and well-known is Expected Goals or xG, this stat measure whether a given shot, depending on the position, will result in a goal. Expected goals give how good a shooting chance was, based on how often similar opportunities have resulted in a goal (Green, 2012). It provides a metric between 0 and 1, with 1 being the most likely to score (Green, 2012). Next to xG, there are quite a few lesser-known ones, such as expected assist or xA, which is the assist equivalent of xG. There are more that will be mentioned later in this thesis.

This research aims to compare prediction models created with advanced statistics and traditional statistics to determine which set of statistics is not only more accurate but also more effective. The research question formulated with this goal in mind reads as follows: *Can advanced statistics be used to build a better performing prediction model for football match results compared to traditional statistics?*

Sub-research questions have been proposed to answer the research question better. These are:

1. How do advanced statistics perform when predicting match results?
2. How does advanced statistics prediction performance compare to the performance of traditional statistics?
3. Which variables are most important when predicting football matches?

The research will be conducted by creating machine learning models that predict future games based on the traditional and advanced statistics of past games, respectively. The results obtained from these models will then be compared on different performance metrics to determine which statistics type performs better.

This thesis is organized as follows: the existing literature on advanced statistics in sports and, more specifically, football will be reviewed, followed by an explanation of the approach and methods chosen for the analysis and finally, the findings will be reported, and the research conclusion will be drawn.

2 Literature Review

The existing literature on predicting football results will be discussed in this chapter to better understand advanced statistics and the proposed research questions.

2.1 Advanced Statistics

Football match analysis has been around for a long time, and Charles Reep is often regarded as the pioneer of the field (Pollard, 2002). Reep started tracking match statistics after World War II using pencil and paper (Pollard, 2002). His analysis helped him create the long ball strategy many teams have used when he concluded that getting the ball forward as fast as possible leads to more success (Larson, 2001). After Reep, the analysis of matches became more prevalent, leading to teams creating competitive advantages over their competitors based on the data they had collected (Pollard, 2002). The advancement of technology and the analysis of matches led to advanced metrics in different sports (Holman, 2018). Advanced metrics have their roots in sabermetrics, which Bill James created for Major League Baseball. He described it as the "search for objective truth in baseball through statistical analysis" (Hirsch & Hirsch, 2014). Advanced statistical approaches were used to look deeper than traditional box scores statistics (Giasemidis, 2020). Advanced analytics has been adopted by many sports, including football. However, the adoption of it by football has been slower compared to other sports, like American football and baseball (Fernandez et al.,2021). Yet, advanced statistics are becoming increasingly popular in strategy, decision-making, and player performance indication (Morgulev et al., 2018).

Keeping up with the large amount of data that could be collected by watching and analyzing a game nowadays would be very difficult to do with just pen and paper, even though that is how it originated. Therefore, large sports data companies, like Opta sports and Statsbomb, employ analysts who watch and collect data of every game, which is made possible by the many cameras and trackers placed in and around football pitches nowadays to make further analysis (*Opta Sports*, n.d.). Each game is watched by two analysts, one for each team and every pass and touch on the ball is recorded by these analysts and saved in a central database where it can be found by teams, fans, websites, and television channels. Matches are also re-watched after they are finished by a different set of analysts to make sure every event that was recorded was done so correctly. Optasports can thus use further calculations and older data to create a more nuanced perspective for the events of a match. The definitions of the advanced statistics used in this thesis were obtained from Opta sports (Opta Event Definitions, n.d.) and these are:

Table 1: Advanced Statistics (Opta Event Definitions,n.d)

Advanced Statistic	Explanation
Expected Goals (xG)	“measures the quality of a shot based on several variables such as assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance. Adding up a player or team’s expected goals can give us an indication of how many goals a player or team should have scored on average, given the shots they have taken.”
Non-Penalty Expected Goals (NPxG)	This is the same as Expected goals but taking penalties, which have an xG of almost 1 out of the equation.
Expected Assists (xA)	“measures the likelihood that a completed pass will become a goal assist. It considers several factors including the type of pass and end-point and length of the pass. Adding up a player or team’s expected assists gives us an indication of how many assists a player of the team should have had based on their build-up and attacking play.”
Key Passes	“The final pass leading to the recipient of the ball having an attempt at goal without scoring.”
Pressures	“Number of times applying pressure to opposing player who is receiving, carrying or releasing the ball.”
Progressive carries	“Carries that move the ball towards the opponent’s goal at least 5 yards or any carry into the penalty area.”
Goal or Shot-creating Actions (GCA or SCA)	“The two offensive actions directly leading to a goal or shot, such as passes, dribbles and drawing fouls.”

The reason for this more nuanced perspective of statistics is simple. For example, two passes can be notated precisely the same but can have a completely different meaning in the game's scope. A pass backward to a teammate standing nearby is much less indicative of performance than a pass forward to a teammate standing in a dangerous position, even if the teammate does not score a goal (‘Data Analytics in Soccer’, 2021). Another example is two teams with the same number of goals scored, but one has much lower expected goals per game. Using the knowledge brought by the xG, it can be decided that one team is much better and the other one just might be lucky (Luzum & Model, 2022).

Furthermore, one more advanced statistic used in this thesis was the Soccer Power Index or SPI. The SPI is the percentage of the available points a team is expected to get from each match (Boice, 2018). SPI is based on three pillars, the market value of the team according to Transfermarkt.com, the offensive strength of the team, or how many goals it is expected to score and the defense strength of the team, or how many goals it is expected to concede, these are calculated using the Expected Goals for each team in a match (Boice, 2018). This index is calculated again for each new season based on the last rating of the previous season for a team and the difference in market value compared to the previous season (Boice, 2018). SPI is also adjusted after every match, depending on prior performances.

2.2 Prediction literature

As mentioned before, football is arguably the world’s most popular sport. Therefore, there is no shortage of literature surrounding the prediction of its matches. Predicting match outcomes based on game statistics has become much more popular because more data is available for public use (Wheatcroft, 2020). Many studies have researched the prediction of soccer matches using vastly different types of methodologies. Most of the research into match outcome predictions is performed by gambling organizations to aid oddsmakers (Ulmer et al., 2013). However, attempts at predicting match results are not new, as there have been analyses dating back to 1956 (Moroney, 1956). Despite this, it was not until 1974 that it was proven that match results are not just based on luck and can be predicted depending on the variable choices (Hill, 1974).

Forecasting football match results can be done in two different ways; first, there is the result-based approach which models the probability of a game being a Home Win, an away win, or a draw, and second the goal-based approach, which models the number of goals scored for home and away teams (Goddard, 2005). Goal-based studies have been done longer and have also been tried and tested much more than result-based. However, neither approach differs drastically in forecasting results (Goddard, 2005). This, however, does not stop the debate about which approach is better, as it is still happening now (Egidi & Torelli, 2020). Many different methods have been tried to achieve better accuracy when predicting match results, mainly statistical techniques, or machine learning (Anfilets et al., 2020). A common practice when forecasting match results is following a Poisson distribution, treating the goals for each team as separate variables, originating with Maher (1982) and Dixon and Coles (1997). Maher (1982) claimed in his groundbreaking paper that the number of goals scored by either team is independent of each other and that the number of goals is based on the performance of the attack and defense; therefore, to predict the outcomes, he gave separate scores for the home and away attack and defense. Match results depend on many reasons, so predicting the exact result of matches is nearly impossible. (Arabzad et al., 2014). The low number of goals in football, compared to other sports, causes the probability of a surprising result to increase because even the best teams can have a bad or unlucky day (Beal et al., 2020).

Machine learning has been used more and more for predicting football results. Prasetio & Harlili (2016) reached an accuracy of 69.5% using logistic regression but leaving games that ended as draws out of it, only modeling Home Wins or away wins. The teams' defenses were the essential variables in this research (Prasetio & Harlili, 2016). Baboota & Kaur (2018) used many different machine learning techniques, but the ones that returned the highest accuracy were Gradient Boosting and Random Forest. Both achieved an accuracy score of 57%, with Gradient Boosting being better at modeling games that ended in a draw. Pugsee and Pattawong (2019) also used Random Forest to predict results. Using four seasons of premier league data, they reached an accuracy of 80% using random forest (Pugsee & Pattawong, 2019).

On the other hand, Artificial Neural Networks are one of the most applied methods for forecasting sports results when considering machine learning (Bunker & Thabtah, 2019). For example, Guan & Wang (2021) reached forecasting accuracy above 70% with Artificial Neural Networks and Rudrapal et al. (2021) used a multilayer perceptron and attained an accuracy of 73.57%. Kundu et al. (2021) achieved an accuracy of 58.8% using a type of Support Vector Machines named sequential minimal optimization, using 12 years of historical data of the Premier League. Using 12 years of data, in this case, worked against the model because over 12 years, the data became much more random due to varying team strengths, clubs getting new owners and having more money at their disposal and football adapting to popular tactics and continuing to evolve. This would then, in turn, increase the difficulty of predicting match outcomes (Kundu et al., 2021).

Across these studies, there have also been a lot of variables used to predict the outcomes, with differing results. Most have used game-day data such as the number of goals, yellow and red cards, possessions, or the number of shots, otherwise known as traditional statistics. Researchers have also attempted to include the recent performances of the teams in question as a variable to aid in the prediction, both using the name “form” (Hucaljuk & Rakipović, 2011; Ulmer & Fernandez, 2013). Baio & Blangiardo (2010) also incorporated “home-field advantage” as a feature in their prediction model. Others have also tried predicting using ratings given in the popular video game FIFA by EA Sports (Baboota & Kaur, 2019, Arntzen & Hvattum, 2021). Partida et al. (2021) used the advanced statistic xG in their predictive model, but they made an alteration by adding home-field advantage to the statistic. This resulted in accuracies on par with Las Vegas oddsmakers; the accuracy was around 70%, but they did not use machine learning techniques (Partida et al., 2021). Other than this, the use of advanced statistics in predicting match results has been lacking in football. Advanced statistics have been used more in studies concerning hockey and basketball. Weissbock et al. (2013) used both advanced and traditional statistics to predict hockey games in the National Hockey League (NHL). They concluded that traditional statistics are better for single games and advanced statistics are better across a whole season. Morrison & Rad (2018) used advanced stats combined with traditional statistics to predict NHL matches and concluded that 5 of the 6 most important features in the dataset were advanced statistics. Wang & Fan (2021) compared the predictive abilities of both traditional and advanced statistics for basketball matches in the National Basketball Association (NBA). They concluded that advanced statistics outperform traditional ones in the context of the NBA (Wang & Fan, 2021).

This closer look at the literature shows that only a little work in the literature uses advanced statistics and that more work is necessary for football regarding the predictive abilities of advanced statistics. This thesis will attempt to determine the predictive capabilities of advanced statistics in football to fill this literature gap. Also, comparing it to the capabilities of traditional statistics.

3 Data Collection

To achieve the goals set at the beginning, relevant data is needed. Many companies have been collecting football data lately, but not all have followed all the advanced statistics closely or made them accessible to the public.

The data set used in this research was scraped from fbref.com, a website devoted to tracking statistics for football teams and players, using SelectorGadget ("All About FBref.Com", n.d). Fbref falls under the Sports-Reference umbrella, which also has websites dedicated to other sports such as baseball, basketball, and American football. Fbref has also begun to track advanced statistics per game due to the rise in usage and popularity in partnership with Statsbomb (*"XG Explained"*, n.d). Fbref contains the advanced statistics for the top 5 leagues in Europe starting from the 2017-18 season until the most recent 2020-21 season. The top 5 leagues are the five leagues with the highest country coefficient according to UEFA ("UEFA.com", n.d.). These leagues are the Spanish La Liga, the English Premier League, the Italian Serie A, the French Ligue 1, and the German Bundesliga. The Spanish, English, Italian and French leagues contain 20 teams, and each team plays 38 games which amounts to 380 total games per season per league. The German Bundesliga has only 18 teams, and each team plays 34 games a season which equals 306 games a season. At the beginning of March 2020, all the European League's seasons were put on hold due to the COVID-19 pandemic; eventually, all except the French League 1 resumed, which resulted in the French League 1 only playing 279 games in the 2019-20 season. Some seasons also contain fewer games that could be used due to some teams beginning their seasons later, for example, starting on matchday two instead of 1 and then playing the missed game later in the season. Three separate datasets were created, each with different "burn-in" periods (Wheatcroft & Sienkiewicz, 2021). A "burn-in" period is a certain number of games that are not taken into the model to allow it to learn about each team, find out what they are good and less good at, and learn tendencies in the leagues and data (Wheatcroft & Sienkiewicz, 2021). It can be seen as the number of games the average of the statistics are taken from. The first "burn-in" period was 1 game, the second was 3 games and lastly, 5 games were used as the "burn-in" period. This means a burn-in period of 1 game uses only the previous game as a reference. A burn-in period of 3 and 5 games takes the average of each statistic over the past 3 and 5 games respectively. This leads to more accurate data but slightly fewer games to predict. The final dataset consisted of 6993 total games when 1 game was used as a “burn-in” period, 6602 games when 3 games were used as a “burn-in” period and 6212 games when 5 games were used as a “burn-in” period.

For the Bundesliga and the covid-shortened 2019-20 Ligue 1 season, the first 200 games are used as training and the rest as testing and for the other leagues, the first 300 games are used as training and the rest as testing. This results in 5500 games in the training data and 1493, 1102, and 712 games in the testing datasets, respectively. The data sets contain 59 columns which are the statistics for both Home and Away teams. These statistics can be seen in Table 1. The final column is the result of the match, which can be “Home Win”, “Draw” or “Away Win”. The data set is split even further into the advanced data set and the traditional data set. The advanced dataset contains 27 columns, the 26 predictor variables and the result. The traditional data set contains 35 columns, 33 predictors and the result. There are fewer columns in these two data sets because the names of the teams and the score result of the match being predicted are removed from further analysis. The research will be result-based, which is why the score of each match is transformed to either “Home Win”, “Away Win” or “Draw”.

Table 2: Variables used in the analysis

Variable	**Description (for both home and away teams)**	**Name in the data set**	**Dataset**
Team	Team names	HomeTeam/AwayTeam	Both
Goals	Average of goals scored per game	HomeGoalsMean/AwayGoalsMean	Traditional
xG	Average xG per game	HomexG/AwayxG	Advanced
NPxG	Average NPxG per game	HomeNPxG/AwayNPxG	Advanced
xA	Average xA per game	HomexA/AwayxA	Advanced
SCA	Average SCA per game	HomeSCA/AwaySCA	Advanced
GCA	Average GCA per game	HomeGCA/AwayGCA	Advanced
Key Passes	The average number of key passes per game	HomeKeyPasses/AwayKeyPasses	Advanced
Progressive Passes	The average number of progressive passes per game	HomeProgressivePasses/AwayProgressivePasses	Advanced
Pressures	The average number of pressures per game	HomePressures/AwayPressures	Advanced
Successful Pressures	The average number of successful pressures	HomeSuccesfulPressures/AwaySuccesfulPressures	Advanced
Successful Pressures Percentage	The average percentage of the number of successful pressures per game	HomeSuccesfulPressurespercentage/AwaySuccesfulPressurespercentage	Advanced
Progressive Carries Distance	The average distance of progressive carries per game	HomeProgressiveCarriesDistance/AwayProgressiveCarriesDistance	Advanced
Progressive Carries	The average number of progressive carries per game	HomeProgressiveCarries/AwayProgressiveCarries	Advanced

SPI	Soccer Power index of Home team and Away team	spi1/spi2	Advanced
Shots	The average number of shots per game	HomeShots/AwayShots	Traditional
Shots on Target	The average number of shots on target per game	HomeShotsonTarget/AwayShotsonTarget	Traditional
Yellow Card	The average number of yellow cards per game	HomeYellowCard/AwayYellowCard	Traditional
Red Card	The average number of red cards per game	HomeRedCard/AwayRedCard	Traditional
Touches	The average number of touches per game	HomeTouches/AwayTouches	Traditional
Interceptions	The average number of interceptions per game	HomeInterceptions/AwayInterceptions	Traditional
Blocks	The average number of blocks per game	HomeBlocks/AwayBlocks	Traditional
Pass Completion	Average pass completion percentage per game	HomePassCompletionpercentage/AwayPassCompletionpercentag	Traditional
Crosses	The average number of crosses per game	HomeCrosses/AwayCrosses	Traditional
Corners	The average number of corners per game	HomeCorners/AwayCorners	Traditional
Tackles	The average number of tackles attempted per game	HomeTackles/AwayTackles	Traditional
Tackles Won	The average number of tackles won per game	HomeTacklesWon/AwayTacklesWon	Traditional
Aerial Duels Won	The average number of aerial duels won	HomeAerialDuelsWon/AwayAerialDuelswon	Traditional
Aerial Duels Lost	The average number of aerial duels lost	HomeAerialDuelsLost/AwayAerialDuelsLost	Traditional
Aerial Duels Won Percentage	Average of the percentage of aerial duels won per game	HomeAerialDuelsWonpercentage/AwayAerialDuelsWonpercentage	Traditional
Result (Target)	Home win, Draw or Away win	results	Both

4 Methodology

The classification methods that will be used to predict whether a match results in a Home Win, Away Win or Draw will be explained in this section. These methods are Multinomial Logistic Regression, Random Forests and (Artificial) Neural Networks.

4.1 Multinomial Logistic Regression

To understand multinomial logistic regression, it is better to start with binary logistic regression or just logistic regression. Binary logistic regression is a generalized linear model that tries to estimate the probability of the predictors used belonging to one of the two results, for example, if instead of Home Win, Away Win and Draw, there were only two choices (Starkweather & Moske, 2011). These probabilities will range between 0 and 1, with a cut-off at 0.5 used to pick the result (Starkweather & Moske, 2011). The probability of the statistics belonging to one of the results is calculated by the logistic function which is (James et al., 2013):

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+ e^{\beta_0+\beta_1 X}} \tag{1}$$

This formula gives an S-shaped curve called the "Sigmoid curve" (James et al., 2013). An example can be seen below (James et al., 2013).

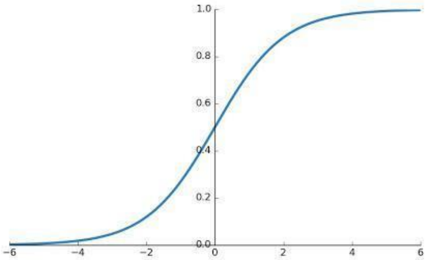

Figure 1: An example of a Sigmoid curve

p(X) is the probability of being a category, in this case, in binary classification, either "Home Win" or "Away Win". The betas (β), the effect each predictor has on the results, are unknown and are therefore estimated using the maximum likelihood function (James et al., 2013). The likelihood function is (James et al., 2013):

$$\ell(\beta_0, \beta_1) \; = \; \prod_{i:y_i=1} p(x_i) \; \prod_{i:y_i=0} (1 - p(x_i)) \tag{2}$$

The betas (β) that are chosen are chosen such that the formula is maximized. The first formula, *p(X),* turns into this formula after some tinkering (James et al., 2013):

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \tag{3}$$

The left side of this formula is called "odds" (James et al., 2013). This side can take any value between 0 and infinity, which is why further tweaking is needed. The logarithm of both sides is taken (James et al., 2013):

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{4}$$

The left side is now called the "log-odds" or "logit" (James et al., 2013). To make predictions, formula (1) must be filled in with the betas you get in formula (4), to obtain the probabilities of each class. Multinomial logistic regression generalizes logistic regression to more than two possible results. In Multinomial Logistic Regression, one of the results gets set as the baseline; in this research "Home Win " was used as the baseline (James et al., 2013). Multinomial Logistic Regression is, in essence K-1 binary logistic regression models; in this research, K = 3 because there are 3 possible classes, therefore 2 logit models are needed. Whether it is a draw if the possible results are draw or home win and whether it is an away win if the possible results are away win and home win, since home win is the reference level. The logit odds are then filled in formulas similar to formula (1). For the baseline (home win) is the probabilities calculated as follows:

$$\hat{p}(Home\ Win) = \frac{1}{1+\sum_{K=2}^{K} e^{\beta_0+\beta_1 X_1+\ \dots+\beta_k X_k}} \tag{5}$$

And for draw or away win it is calculated as follows:

$$\hat{p}(Draw\ or\ Away\ Win) = \frac{e^{\beta_0+\beta_1 X_1+\ \dots+\beta_k X_k}}{1+\sum_{K=2}^{K} e^{\beta_0+\beta_1 X_1+\ \dots+\beta_k X_k}} \tag{6}$$

These models then return three probabilities between 0 and 1 which sum up to 1 total; the highest probability is the overall prediction. Consequently, only one of the classes can have the highest probabilities since they all are dependent on each other. This explanation of Logistic Regression in formula (1) was done when there is only one predictor β_1 but naturally, it is possible to also have more than one predictor, in this thesis, there are 27 and 33 for the advanced and traditional statistics respectively. Hence why the predictors are filled from β_1 to β_{k}.

Before a multinomial logistic regression can be executed, a few criteria must be met. There must be an appropriate outcome type, a linear relationship between log odds and the independent variables, no outliers in the data and lastly no multicollinearity between the predictors (Starkweather & Moske, 2011).

As mentioned before, a multinomial logistic regression with 3 classes is essentially two logistic regressions and this also means two betas for every predictor variable. The absolute value of these two betas is used to determine which predictor has the most impact on the result of this model, known as variable importance (Kuhn, 2008).

4.2 Random forests

Random forests start with Decision Trees, these are classifiers that use a tree-like structure to model the relationship between features and the class (James et al, 2013). Random forests can be used for both classification and regression problems (James et al, 2013).

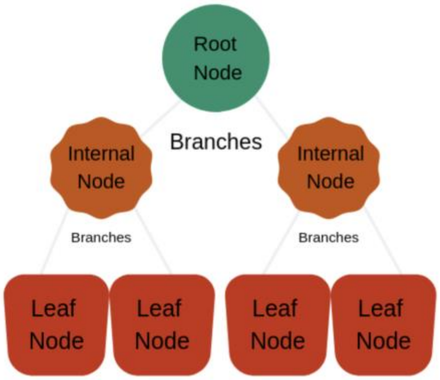

Figure 2: Example of a Decision Tree.

The classification begins at the root node, where the first feature is passed, then the decision node that requires choices based on the feature, these decisions are split into branches that return potential classes and then is the final decision (class) given at the end at the leaf node (Lantz, 2015). A decision tree asks a series of simple questions to make an informed final decision. Random Forests, also known as decision trees forests, make an individual decision trees ensemble. Each of these trees classifies the components, and whichever class gets assigned the most is the model's overall prediction. How many trees need to be created depends on the context and is done on a trial-and-error basis, but generally larger data sets need more trees than smaller data sets (Oshiro et al., 2012). Each separate decision tree uses a different group of predictors to make its classification, which is why Random Forest is well equipped for large data sets (James et al, 2013). This gives other predictors a chance if there is a powerful predictor in the data. Random forests tend to be a bit less interpretable than decision trees and have less correlation between trees. At every split of each tree are $\sqrt{p}$ predictors used to cast a vote for a class, with p being the total number of predictors (James at al., 2013). Random forests also return a so-called "Out-Of-Bag Error Rate" (OOB) because random forests only use a part of all the observations in every tree, they use the ones left out to test the performance of the model (James et al, 2013). Random Forests are not very interpretable, but luckily enough, the algorithm returns the importance of the variables used. The importance of the variables is based on the Gini index. The Gini index indicates from which class the observations come. A high Gini index means that the observations are not from one class and vice versa. This is also called Node Purity (James et al., 2013). Pmk indicates the percentage of training observations in a tree that belongs to a specific class.

$$\mathrm{G} = \prod_{k=1}^{k} \hat{p}mk\ (1 - \hat{p}mk) \tag{7}$$

Variable importances are also obtainable through Random Forests. In this research, the mean decrease accuracy is used to measure importance. In essence, it is how much the model's accuracy would drop if a certain predictor were to be removed from it (Hoare, 2018). A single predictor, in every tree, is randomized while keeping the others constant and the difference in OOB is calculated. The mean is taken across every tree and normalized to obtain the mean decrease accuracy (Breiman, 2001 & Hoare, 2018). This is done for every variable to obtain each predictor's importance score.

4.3 Neural Network

Artificial Neural Network (ANN) mimics the workings of a biological brain when it uses artificial neurons to solve a problem. In a human brain, incoming signals are received by a cell's dendrites. This signal is weighted by its importance or frequency and then send down the axon, and at the axon's terminal is, this signal processed again and then passed again to other neurons, how this works can be seen in figure 3 below (Lantz, 2015).

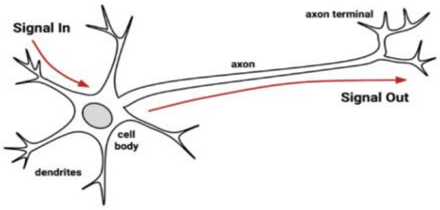

Figure 3: Biological Neural Network

An artificial neural network has a similar working; each input variable is weighted by importance and then passed to the cell's body where they are weighted again. After this, they are passed to an activation function in a cell's body, which then passes the output signal, which is the class. As can be seen below in figure 4 (Lantz, 2015). *X* are the input variables, *w* the weights *f* is the activation function and *y* the final output.

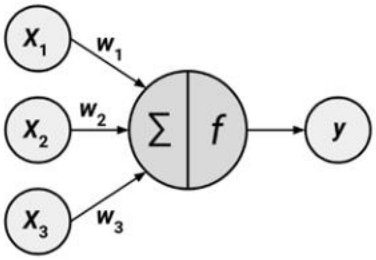

Figure 4: Simple Neural Network.

There are many different neural networks, but they all can be differentiated by their activation function, network topology and training algorithm (Lantz, 2015). The activation function is the mechanism inside the network that transforms the input variables to the eventual outcome, every layer except the input layer contains an activation function (Lantz, 2015). The topology is the number of neurons and layers a neural network contains; this also determines how extensive a problem can be solved with the ANN (Lantz, 2015). More layers and neurons lead to more extensive problems and more computational time. Finally, the training algorithm specifies how the weights of the input signal are calculated, these weights in essence, give a ranking to the importance of the predictor variables and the most used algorithm is backpropagation (Lantz, 2015). This entails that the starting weights are set randomly at the beginning, the algorithm goes through the neural network from the input layer to the output layer to get a result, this result is then compared to the actual result, this difference is then used to adjust the starting weights (Lantz, 2015). The learning rate adjusts the starting weights; this number can be chosen before creating the model (Brownlee, 2019). Each time this process is done is called an epoch (Brownlee, 2019). The more epochs are done, the better, but it also takes more time.

Now that this is clear, can the ANN used in this research be explained a bit better. The input (first) layer has as many neurons as there are predictors in the data, and the output layer contains, in the case of classification, as many nodes as there are classes, the only layer where the nodes can be chosen freely are the hidden layers (Lantz, 2015). Selecting the least number of nodes with a good result is best since more nodes add more computational time. There are many activation functions to choose from, but the one used in this research, since it is a multiclass classification problem, is the SoftMax activation function (Saxena, 2021). The SoftMax activation function returns the probability of the statistics belonging to each of the three classes in this research (Saxena, 2021).

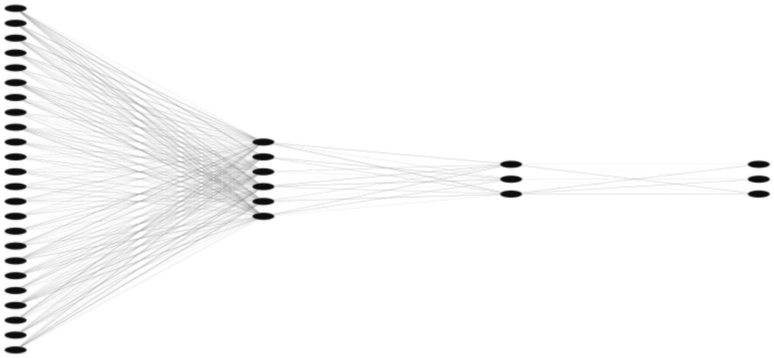

Figure 5: Example of a multilayer Neural Network (Lantz, 2015).

Figure 5 is an example of a neural network that could be used for the advanced statistics data set. With the 24 input neurons, 6 and 3 neurons in the hidden layers and the three neurons for the output, which stand for, one for each of “Home Win”, “Draw” and “Away Win”.

Variable importance scores are calculated similarly to Random Forest for Neural Network. Individual predictors are randomized while others are held constant and the drop in prediction accuracy is seen as the variable importance (Breiman, 2001 & Molnar, 2020).

4.4 Model Performance Measures

Each model will be evaluated on 3 performance metrics: their Accuracy, F-measures and Ranked Probability Scores (RPS). Accuracy is calculated with the following formula (Lantz, 2015):

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \tag{8}$$

A higher accuracy naturally means more correct predictions and a better-performing model. However, accuracy does not always tell the whole story of the model.

The F-measure creates a global result of the True Positives, True Negatives, False Positives and False Negatives of a classifier's predictions. Predictions for each class are each given one of these labels when they are assigned to a class (Lantz, 2015)

- True Positive: Correctly Classified as the class;

- True Negative: Correctly classified as another class;
- False Positive: Incorrectly Classified as the class;
- False Negative: Incorrectly classified as another class.

The F-measure is subsequently calculated, per class, using these labels. With the following formula (Lantz, 2015):

$$F - Measure : \frac{2\ x\ True\ Positive}{2\ x\ True\ Positive + False\ Positive + False\ Negative} \tag{9}$$

The F-measure for each class is subsequently summed up and divided by the number of classes, 3, to obtain the global F-measure of the model (Shmueli, 2019).

The final metric, the Ranked Probability Score (RPS), evaluates the model deeper. The RPS is a strictly proper scoring rule (Murphy, 1969). The RPS assesses the probabilities given to each result by the model to evaluate the model instead of strictly looking at the highest probabilities. According to Constantinou and Fenton (2012), it is the most adequate scoring rule to evaluate match forecasts. Constantinou and Fenton (2012) argued that the results of a football match are on an “ordinal scale”, meaning a home win is closer to a draw than an away win, and the scoring rule should be chosen with this in mind. The RPS meets this criterion and is calculated as follows (Murphy, 1969 & Constantinou & Fenton, 2012):

$$RPS\ = \frac{1}{r-1}\sum_{i=1}^{r}\left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j\right)^2 \tag{10}$$

Where *r* are the potential outcomes (Home Win, Draw and Away Win), *p* is a vector of the predicted probabilities and *e* is a vector of actual probabilities (Murphy, 1969). J shows which components in the vectors are being referred to. *i* refers to the number of probabilities in the vector, which is the same number as the potential outcomes. RPS calculates the distance between the predicted probability and the actual probability, the lower this result is the better the model is at predicting the results of matches. But a large discrepancy in forecasted probabilities also results in a bigger penalty.

5 Results

Before the models were trained, the correlations between the statistics and the results were checked. The correlations with the result “Draw” were extremely low compared to the other results, which implies that predicting draws will be tough for these models, these correlations can be seen below in tables 2 and 3. This will be even more difficult because most of the results are a win for either the home or away team. Over 70% of each dataset results in either a win for the home team or the away team.

Table 2: Highest 5 features in correlation with each result in advanced dataset

Home Win	Draw	Away Win
Home Win = 1.000	Draw = 1.000	Away Win = 1.000
SPI1 = 0.257	HomePressures = -0.006	SPI2 = 0.262
HomeProgressiveCarries = 0.228	AwayPressures = -0.010	AwayProgressiveCarries = 0.241
HomeProgressivePasses = 0.223	AwaySuccesfulPressuresPercentage = -0.015	AwayProgressivePasses = 0.241
HomeXA = 0.212	AwaySuccesfulPressures = -0.021	Away NPxG = 0.238

Table 3: Highest 5 features in correlation with each result in the traditional dataset

Home Win	Draw	Away Win
Home Win = 1.000	Draw = 1.000	Away Win = 1.000
HomeTouches = 0.220	HomeYellowCard = 0.034	AwayTouches = 0.250
HomeShots = 0.200	AwayBlocks = 0.026	AwayPassCompletionPercentage = 0.213
HomeShotsonTarget = 0.197	HomeRedCard = 0.016	AwayShotsonTarget = 0.212
HomeGoalsMean = 0.190	HomeAerialDuelsLost = 0.009	AwayShots = 0.208

The previously discussed classification techniques will be shown and analyzed in the remainder of this section. The obtained results can be seen in the table below. These results were achieved after the models were trained with different hyperparameters to obtain optimal results. For each method is the model and result with the highest accuracy explained further with confusion matrices and partial dependence plots (PDP). The highest accuracies were always reached when using 5 games as burn-in period, similar to Wheatcroft & Sienkiewicz (2021). The models are ordered from the lowest RPS to the highest in table 4. The advanced dataset models obtained a mean accuracy of 52.93%, a mean F1-measure of 0.400 and a mean RPS of 0.2087, while the traditional datasets obtained a mean accuracy of 51.89%, a mean F-measure of 0.391 and a mean RPS of 0.2129. Furthermore, multinomial logistic regression delivers the highest accuracies and lower RPS. While Random Forests deliver the highest F-measure

Table 4: Results of models trained

Method	Burn-in	Dataset	Accuracy	F-Measure	RPS
Multinomial Logistic Regression	1 Game	Advanced	52.51%	0.388	**0.2021**
Multinomial Logistic Regression	1 Game	Traditional	51.98%	0.377	0.2047
Neural Network	1 Game	Traditional	51.64%	0.379	0.2070
Random Forest	1 Game	Traditional	52.04%	0.422	0.2072
Neural Network	1 Game	Advanced	52.24%	0.380	0.2081
Random Forest	1 Game	Advanced	51.31%	0.418	0.2083
Multinomial Logistic Regression	3 Games	Advanced	52.72%	0.393	0.2084
Neural Network	3 Games	Advanced	52.63%	0.382	0.2092
Multinomial Logistic Regression	5 Games	Advanced	**54.78%**	0.397	0.2098
Random Forest	3 Games	Advanced	52.27%	0.424	0.2100
Neural Network	5 Games	Advanced	54.78%	0.389	0.2106
Random Forest	5 Games	Advanced	53.09%	0.432	0.2120
Random Forest	5 Games	Traditional	53.79%	**0.429**	0.2138
Multinomial Logistic Regression	5 Games	Traditional	53.23%	0.401	0.2144
Multinomial Logistic Regression	3 Games	Traditional	50.36%	0.381	0.2167
Random Forest	3 Games	Traditional	50.64%	0.389	0.2169
Neural Network	3 Games	Traditional	50.36%	0.388	0.2171
Neural Network	5 Games	Traditional	52.95%	0.357	0.2186

5.1 Multinomial Logistic Regression

The multinomial logistic regression returned maximum accuracies of 54.78% and 53.23% for the advanced and traditional statistics, respectively. The accuracy reached for the advanced statistics is also the highest obtained across all the models. The advanced statistics model had a F-measure of 0.398 and a RPS of 0.2098. The advanced statistics also achieved the lowest RPS with multinomial logistic regression at 0.2021, when using 1 game as burn-in period. While the traditional statistics had a F-measure of 0.401 and a RPS of 0.2144. The confusion matrices for these models can be seen below in tables 5 and 6. The higher F-measure for the traditional statistics can be explained because the model with traditional statistics is better at forecasting draws. Overall, advanced statistics had a mean accuracy of 53.34, mean F-measure of 0.393 and a mean RPS of 0.2067 across the multinomial logistic regression models. While the traditional statistics had a mean accuracy of 50.99%, mean F-measure of 0.386 and a mean RPS of 0.2120 across the models.

Table 5: Confusion Matrix for advanced statistics (5 games)

	Predicted Category		
	Home Win	Draw	Away Win
Home Win	256	4	73
Draw	94	0	52
Away Win	95	3	135

Table 6: Confusion Matrix for traditional statistics (5 games)

	Predicted category		
	Home Win	Draw	Away Win
Home Win	256	6	71
Draw	102	1	43
Away Win	108	9	116

The Multinomial Logistic Regression was done with Home Win as the reference. This implies that 2 models are created, one with Draw relative to Home Win and one with Away Win relative to Home win. This means that every estimate change is also relative to the reference, forecasting a Draw or Away Win over a Home Win. The data sets were checked for multicollinearity before being fed to the models. The advanced model was trained with only 6 variables after removing the other 21 features due to multicollinearity. The traditional data set was trained with 28 features, removing only 4 due to multicollinearity.

The variable importance plots for the models can be seen below in figure 7. It can be seen that for advanced statistics, *spi1* and *spi2* are the two most important variables for predicting matches. As for the traditional statistics, *HomeGoalsMean* is the most important variable. Surprisingly it is *AwayRedCard* and *HomeRedCard* that follow as the next most important variables.

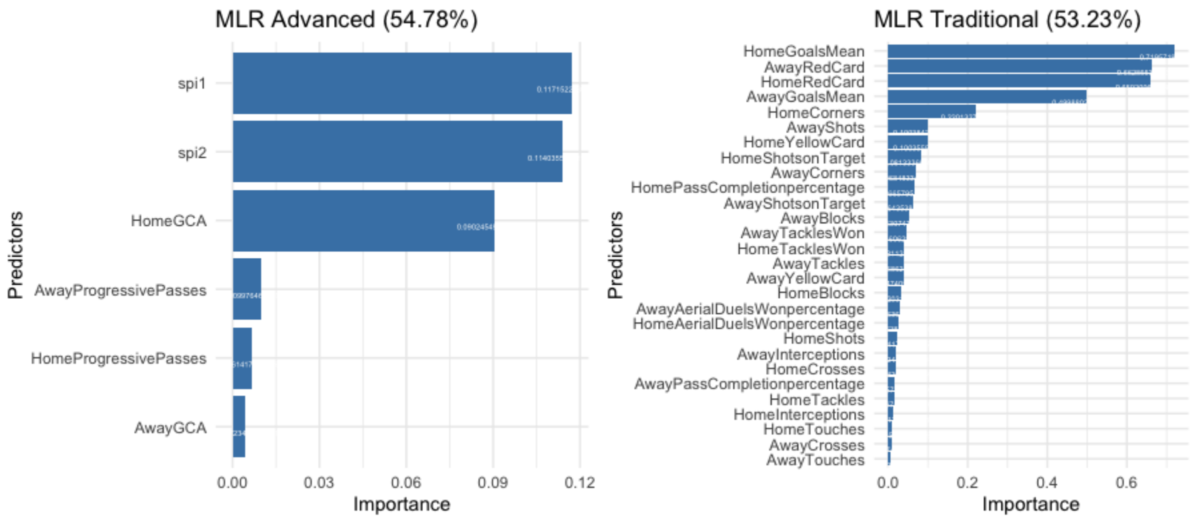

Figure 7: Variable Importance Plots for multinomial logistic regression models

5.2 Random Forest

The second classifier used was Random Forest. First, a default random forest was trained with 1000 trees and $\sqrt{predictors}$ per tree. This returned a result of 51.54% for the advanced statistics and 50.97% for the traditional statistics. After some fine-tuning, the optimal number of trees and variables used at each split was found. For the advanced statistics, it turned out that 1000 trees and 4 variables per split were optimal. While the Out-Of-Bag error (OOB) remained constant after 500 trees for the traditional dataset, so 500 trees were chosen. The number of variables was also optimal at $\sqrt{predictors}$ per split. The confusion matrices for the maximum accuracy of the two datasets can be seen below in tables 7 and 8. It is clear to see that random forests are much better at classifying draws than multinomial logistic regressions and neural networks. The maximum accuracy reached for random forest were 53.09% and 53.79% for the advanced and traditional statistics respectively, both when using 5 games as burn-in period. The advanced statistics model had an F-measure of 0.432 and an RPS of 0.2120. The traditional statistics model had an F-measure of 0.429 and an RPS of 0.2138. The F-measure for the traditional statistics was also the highest one achieved with these data sets. The mean accuracies across the models were 52.22% and 52.15% for the advanced and traditional statistics respectively. While the means for the F-measure were 0.425 and 0.413. And means of 0.210 and 0.212 for the RPS. F-measures for the random forests are generally higher than the other models because random forests are better at forecasting draws than the other models. Which can also be seen in the confusion matrices below (tables 7 and 8).

Table 7: Confusion Matrix for advanced statistics (5 Games)

	Predicted Category		
	Home Win	Draw	Away Win
Home Win	250	85	101
Draw	20	15	19
Away Win	63	46	113

Table 8: Confusion Matrix for traditional statistics (5 Games)

	Predicted Category		
	Home Win	Draw	Away Win
Home Win	257	101	111
Draw	6	12	8
Away Win	70	33	114

The next step in the analysis was the importance of the variables used in the Random Forest. The variable importances from both datasets can be seen below in figure 8. Removing these variables would cause the highest accuracy decrease starting from top to bottom. Both datasets contain highly important variables: *spi* in advanced statistics and *touches* in the traditional statistics dataset.

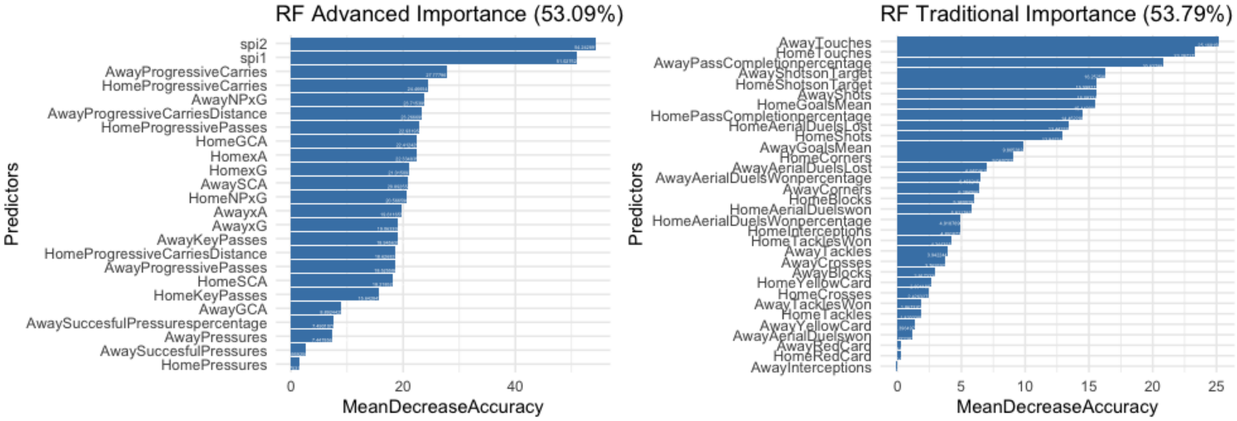

Figure 8: Variable importance for random forests models

Next from the analysis are the partial dependence plots or PDP. These plots can show how a change in a specific predictor variable affects each possible outcome. The PDP in figure 9 below show the effect of *spi2* and *AwayTouches* on the chances of a Home Win because these are by far the most important variables in each dataset. The PDP for draws is a bit harder to interpret, this is the result of draws being harder to predict. These plots make it clear that the two variables have comparable effects on the chances of a Home Win, which could be expected. The vertical axis of these plots shows the chance of the result being a Home Win and the horizontal axis shows what the *spi* is for the away team and the number of touches the away team had in the previous game. The effect of *spi2* on Home Win is stronger than *AwayTouches*, this can also be seen on the vertical axis because the probabilities for *spi2* are between -0.2 and 0.6, while the ones for *AwayTouches* are between 0.1 and 0.4. The marks on the horizontal axis show the distribution of the data points for these variables. For *spi2 the* most data points observed were between 50 and 90, and for *AwayTouches* they were between 500 and 800. The estimated effect is less reliable outside of these boundaries because there were not enough data points to work with. The data points also show where most teams are regarding these statistics.

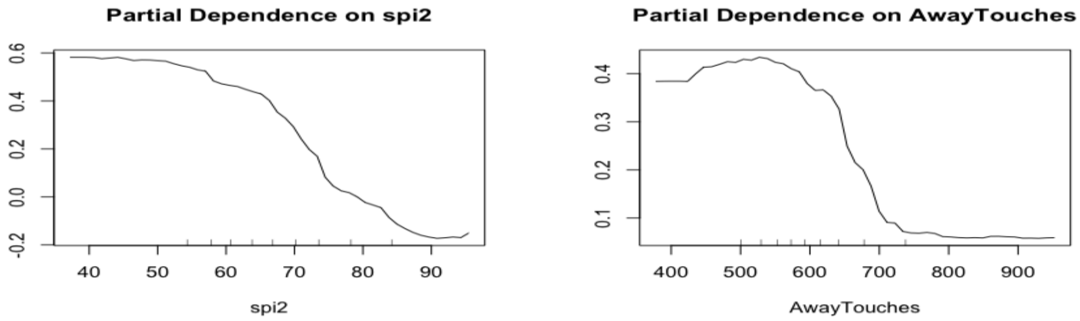

Figure 9: PDPs on Home Win

The plots for the third most important variables of each data set can be seen below in figure 10. These plots are less straightforward, but some conclusions can be taken from them. When the number of *AwayProgressiveCarries* or *AwayPassCompletionpercentage* increases the chances of an Away Win go up. But the probabilities are lower compared to the most important variables shown above, naturally because they are less important and would affect a game less. The lines in these plots also contain interesting deviations, for example, a short decrease for progressive passes between 20 and 40. While between 40 and 60 it shows a drastic increase in the probability of an away win and then levels off again. *AwayPassCompletionpercentage* shows a steady climb on the probability of an away win and then also levels off. The data points for *AwayProgressiveCarries* mainly were between 20 and 60 and for *AwayPassCompletionpercentage* they mainly were between 70 and 85.

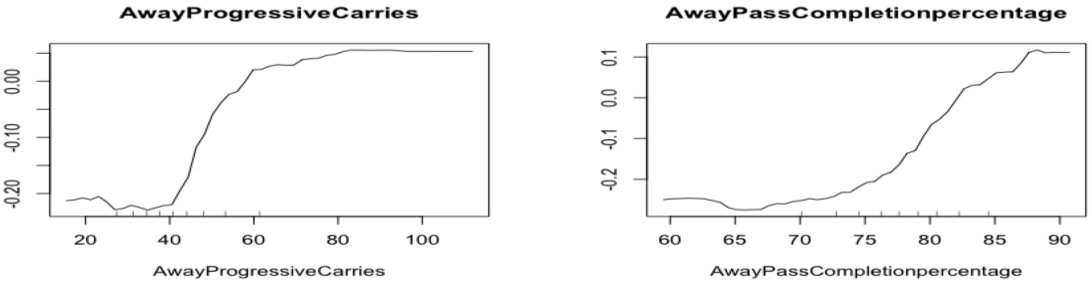

Figure 10: PDPs on Away Win

5.3 Neural Network

Before training the models and tuning the parameters, the variables were normalized, and the results transformed using one-hot encoding. Subsequently, the default neural networks were trained for each dataset used. These default neural networks contained no hidden layers and had only 10 epochs and a batch size of 32. The highest accuracy of these default networks presented was 51.40% for the advanced dataset and 50.77% for the traditional dataset when using only 1 game as the burn-in period. The models were modified heavily next with the tuning of the parameters, this included increasing the number of units per layer, adding more layers, increasing the batch size, and adding more epochs. The models were trained repeatedly with 1, 2 and 3 layers, from 32 to 256 units per layer, a batch size of 16, 32, 64 or 128 and between 100 and 500 epochs for each dataset used. The neural networks reached a maximum accuracy of 54.78% with an F-measure of 0.389 and an RPS of 0.2106 for the advanced dataset. The maximum accuracy for the traditional statistics was 52.95% with an F-measure of 0.357 and an RPS of 0.2186. The trained neural network for the advanced dataset contained 3 hidden layers, 250 epochs and a batch size of 16. While the trained neural network for the traditional dataset had 3 hidden layers, 250 epochs and a batch size of 32. The confusion matrices for the neural networks with the maximum accuracy can be seen below in tables 9 and 10. The neural networks for the advanced statistics had a mean accuracy of 53.22%, a mean F-measure of 0.384 and a mean RPS of 0.2093. Meanwhile the neural networks for the traditional statistics had a mean accuracy of 51.65%, a mean F-measure of 0.375 and a mean RPS of 0.2142. Neural networks however are poor at forecasting draws and have not even predicted one draw correctly. Which can be seen in the confusion matrices below (tables 9 and 10).

Table 9: Confusion matrices for the neural network (5 Games)

Predicted Category			
	Home Win	Draw	Away Win
Home Win	283	109	126
Draw	0	0	0
Away Win	50	37	107

Table 10: Confusion matrices for the neural network (5 Games)

Predicted category			
	Home Win	Draw	Away Win
Home Win	304	128	160
Draw	0	0	0
Away Win	29	18	73

The two models were further analyzed after they were fully trained. The variable importance scores were extracted first from the models, these can be seen below, in figure 11. The variables are on a log scale to highlight the differences more clearly. There are some noticeable trends to be seen in the variable importance plots. The most important variables are like the ones in the Random Forest above with *spi2* and *AwayTouches* leading the way for the two datasets respectively. *HomeTouches* in the traditional dataset is not the second most important variable for the NN model as it was for the Random Forest model. The least important variables, *HomeProgressiveCarries* and *AwayShots*, were more important in the Random Forest. This shows that the two models utilize features differently from each other even though the most important ones are similar.

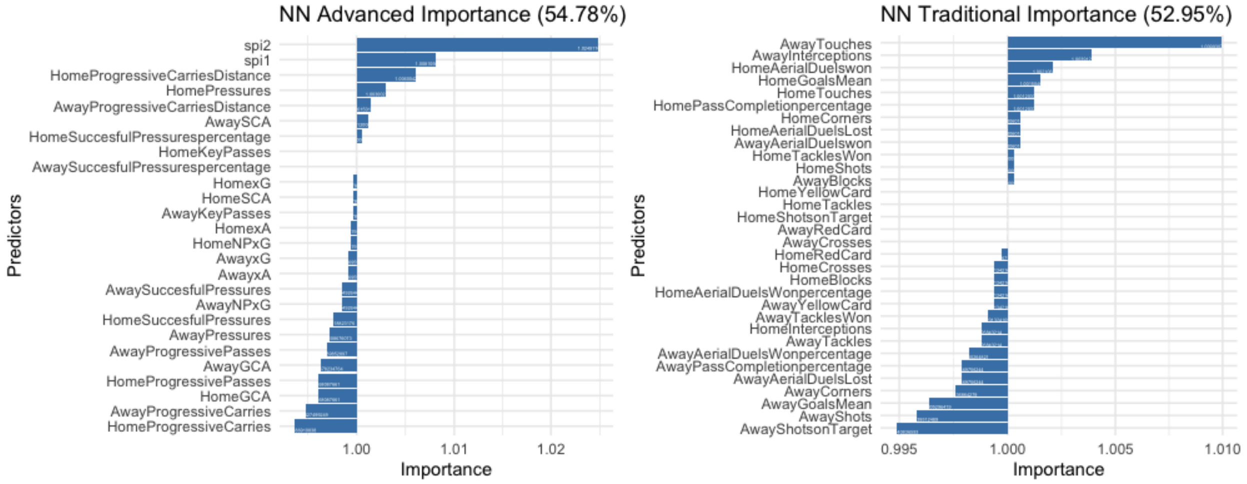

Figure 11: Variable Importance plots for neural network models

The PDPs for the top two most important variables in each data set can be seen below in figure 12. Due to the normalization of the features for the neural network model, all the observations (the marks on the horizontal axis) are very close together, between 0 and 1. The partial dependence plots for neural networks show a nearly linear relationship between some traditional statistics of the variables and the results in the advanced statistics. *spi1* and *spi2* have an exactly opposite reaction on the probability of a Home Win, though *spi1* is less linear than *spi2*. But intuitively both are correct because a higher *spi2* leads to a lower chance of a Home Win and vice versa. Regarding the traditional statistics, an increase in *AwayTouches* would lead to a lower probability of a Home Win. And a rise in *HomeAerialDuelswon* would lead to an increase in the probability of a Home Win. Both are also to be expected when looking at the natural progression of a football match. *AwayTouches* do have the largest effect on a Home Win when strictly looking at the probabilities in the PDP.

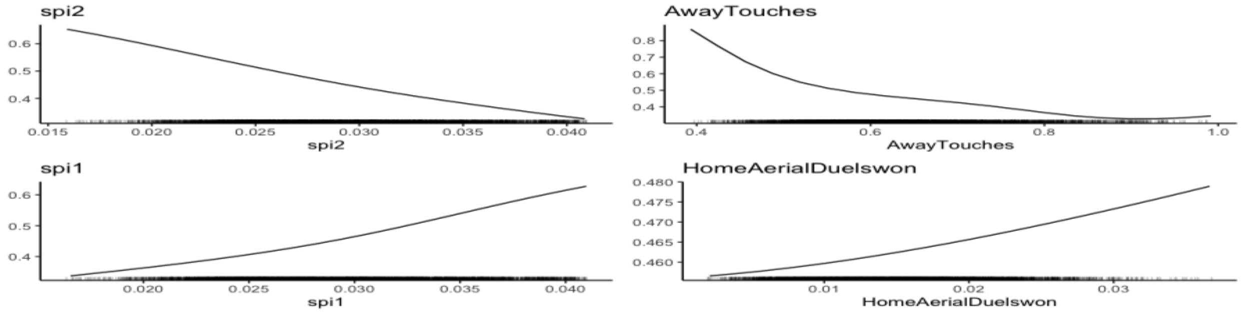

Figure 12: Partial Dependence Plots for “Home Win” for NN

5.4 Variable Importance

As was explained above each method delivers its own variation of variable importance. This is why hybrid datasets with both advanced and traditional statistics using the three burn-in periods were created to be fed to the models. This means 9 models were made, 3 for each method. The datasets contained a total of 59 variables. The models with the highest accuracies for each method were used to determine the variable importance. The burn-in period of 5 games returned the highest accuracy for every model, as was expected. The models returned maximum accuracies of 54.78%, 53.09% and 54.35% for the multinomial logistic regressions, random forests and neural networks respectively. The hybrid models show that combining the two types of statistics does not increase prediction power as none of them outperformed their counterparts in the sections above. The variable importance plots for each method can be seen below in figure 13, with the variables in orange being the advanced statistics and the ones in gray traditional statistics. Only the top 10 variables of each method are shown. The first noticeable thing is that the logistic regression does not consider advanced statistics as important as the rest of the models. While the other models consider them to be the most important statistics. It can also clearly be observed that *spi1* and *spi2* are important predictors in every model used for the advanced statistic. Meanwhile, when it comes to the traditional statistics the variables vary very much depending on the model. For example, extremely important variables in the multinomial logistic regression (*AwayRedCard* and *HomeRedCard*) are not considered important in the other two models. No traditional statistics are considered one of the most important variables of every model, the models all choose different combinations of the traditional statistics.

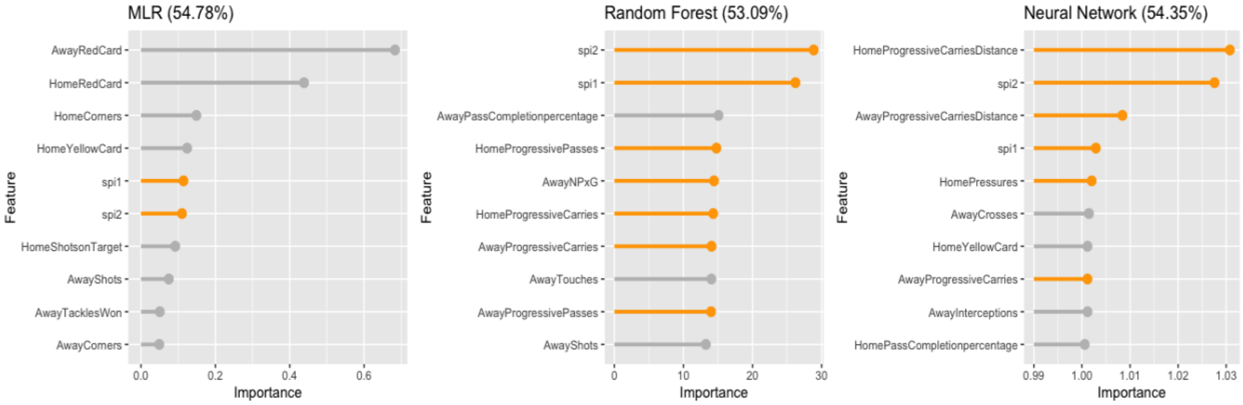

Figure 13: Variable Importance plots for Hybrid Models

6 Conclusion & Discussion

6.1 Conclusion

Predicting something based on previous results is never easy nor simple, especially something with as volatile results as football matches. This difficult task is attempted in this thesis, using advanced statistics while trying to determine if it would lead to differing results compared to traditional statistics. Advanced analytics has recently seen a rise in popularity among sports teams, analysts, and even fans. Many advanced statistics, like xG, xA and Key Passes, are becoming more regular statistics used to evaluate teams and players. The prediction power of this form of statistics and traditional statistics, such as shots on target and the number of corners per match, will be tested in this research. Three different datasets were created for analysis during this thesis and applied to Multinomial Logistic Regressions, Random Forests and Neural Networks. The main research question of this thesis was constructed considering this and it is: *Can advanced statistics be used to build a better performing prediction model for football match results compared to traditional statistics?*

Before answering this question was it imperative to fully grasp what advanced statistics are and what differentiates them from traditional statistics. Traditional statistics started off being tracked using pen and paper and then evolved along with technology to where it is nowadays: multiple analysts tracking the same match through a computer and keeping track of every action in the match. Tracking match events through computers allows analysts to track many more events, which leads to more data points per match, which can be used to put a match into perspective in a more understandable manner. Statistics are “advanced” when they can tell a deeper story about a match than just looking at the box score. Advanced statistics consider how something happened and what it resulted in, and what similar activities have resulted in in the past.

The first two sub-questions are concerning the performance of both the traditional and advanced statistics and which group of statistics is more effective when predicting match results. Looking at the results, presented in the results section above, the conclusion can be drawn that advanced statistics outperform traditional statistics across every performance metric used in this thesis. The mean accuracy for advanced statistics across each method was higher than the traditional statistics’. Each method’s highest accuracy was also reached when using advanced statistics.

Subsequently, the datasets were compared using the F-measure. Advanced statistics perform better again, on average. However, the highest F-measure was reached with traditional statistics.

Lastly, the statistics were compared to each other using Ranked Probability Scores, again, advanced statistics is the better performing model with a lower RPS, on average, compared to the traditional datasets. The lowest overall RPS is also reached using advanced statistics.

The third and final sub-question asks which specific statistics are the most important when predicting match results. Three hybrid datasets were created to be used in the methods and obtain their variable importance scores. From the variable importance plots, it could be seen that the SPI and Progressive Carries from the advanced dataset are variables that were important in each method. Traditional statistics were more divided across the models and no statistic stood out for all the models. But when looking at the variable importance plots from the non-hybrid models it can be seen that the number of Touches for the home and away team appears to be an important predictor for the traditional statistics. Additionally, it can also be concluded that multinomial logistic regression considers different variables more important than random forest and neural networks. While random forests and neural networks are more similar in the predictors they find important.

With the help of the sub-questions and the answers that this research drew from them, it can be concluded that advanced statistics can be used to build a better-performing prediction model for football match results. Advanced statistics returned higher accuracies, F-measure and lower RPS.

6.2 Discussion

This research provides a fresh look into prediction models for match results, but it did not come without its own set of limitations that future research could explore.

From the accuracies provided and compared to previous studies regarding predicting match results, it can be argued that the chosen datasets are a bit general, even though the goal of this thesis was to see how a general dataset of advanced and traditional statistics would compare to each other. It would be interesting to see how different the results would be with a more team-specific dataset. Or with more matchup specific statistics.

Furthermore, many other methods could have been used in this research. Some have proven to be successful in other research. As seen in the Random Forest, the traditional dataset twice achieved higher accuracy than the advanced dataset. It is entirely possible that a different set of methods could have brought different results. Pursuing this further, not many draws were predicted with these methods even though they are an integral part of football; a different method could bring more balance in the prediction accuracy per class.

Lastly, the datasets used in this research contained match information for 5 different leagues, Italian, Spanish, German, French and English, but these leagues are all different and display different playstyles and tactics, which would result in different statistics weighing more in one league compared to another. In the same way, it would be interesting to see this method applied to knock-out competitions, where teams routinely play different styles depending on the opponent they face, as opposed to the course of a whole season where the stronger team usually prevails in the end.

Bibliography

All About FBref.com. (n.d.). FBref.Com. Retrieved 15 May 2021, from https://fbref.com/en/about/

Anfilets, S., Bezobrazov, S., Golovko, V., Sachenko, A., Komar, M., Dolny, R., Kasyanik, V., Bykovyy, P., Mikhno, E., & Osolinskyi, O. (2020). *DEEP MULTILAYER NEURAL NETWORK FOR PREDICTING THE WINNER OF FOOTBALL MATCHES*. 8.

Angelini, G., & Angelis, L. D. (2017). PARX model for football match predictions. Journal of Forecasting, 36(7), 795–807. https://doi.org/10.1002/for.2471

Arabzad, S. M., Araghi, M. E. T., Sadi-Nezhad, S., & Ghofrani, N. (2014). Applied Research on Industrial Engineering. 1(3), 21.

Arntzen, H., & Hvattum, L. M. (2021). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, *21*(5), 449-470.

Baboota, R., & Kaur, H. (2018). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*. https://doi.org/10.1016/j.ijforecast.2018.01.003

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, *37*(2), 253–264. https://doi.org/10.1080/02664760802684177

Beal, R., Middleton, S. E., Norman, T. J., & Ramchurn, S. D. (2020). Combining Machine Learning and Human Experts to Predict Match Outcomes in Football: A Baseline Model. ArXiv:2012.04380 [Cs]. http://arxiv.org/abs/2012.04380

Boice, J. (2018, August 10). How Our Club Soccer Predictions Work. *FiveThirtyEight*. https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brownlee, J. (2019, January 24). Understand the Impact of Learning Rate on Neural Network Performance. *Machine Learning Mastery*. https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Constantinou, A. C., & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, *8*(1).

Data Analytics in Soccer. (2021, June 8). *The Soccer Stands*. https://thesoccerstands.com/data-analytics-in-soccer/

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), 265-280.

Egidi, L., & Torelli, N. (2020). Comparing Goal-Based and Result-Based Approaches in Modelling Football Outcomes. Social Indicators Research. https://doi.org/10.1007/s11205-020-02293-z

Fernández, J., Bornn, L., & Cervone, D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, *110*(6), 1389-1427.

Giasemidis, G. (2020). Descriptive and Predictive Analysis of Euroleague Basketball Games and the Wisdom of Basketball Crowds. *arXiv preprint arXiv:2002.08465*.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. International Journal of Forecasting, 21(2), 331–340.

https://doi.org/10.1016/j.ijforecast.2004.08.002

Green, S. (2012). Assessing The Performance of Premier League Goalscorers. Stats Perform. https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/

Guan, S., & Wang, X. (2021). Optimization analysis of football match prediction model based on neural network. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-05930-x

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, *45*(2), 171-186.

Hill, I. D. (1974). Association Football and Statistical Inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *23*(2), 203–208. https://doi.org/10.2307/2347001

Hirsch, S., & Hirsch, A. (2014). *The Beauty of Short Hops: How Chance and Circumstance Confound the Moneyball Approach to Baseball*. McFarland.

Hoare, J. (2018). How is Variable Importance Calculated for a Random Forest? *Displayr*. https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/

Holman, V. (2018). What is sports analytics. *Agile Sports Analytics*.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299-310.

Hucaljuk, J., & Rakipović, A. (2011). *Predicting football scores using machine learning techniques*. 5.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*

(Vol. 112, p. 18). New York: springer.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, *28*, 1-26.

Kundu, T., Choudhury, A., & Rai, S. (2021). *Predicting English Premier League Matches Using Classification and Regression* (pp. 555–568). https://doi.org/10.1007/978-981-15-5077-5_50

Lantz, B. (2015). *Machine learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R* (1. publ). Packt Publ.

Larrousse, B. (2019). Improving decision making for shots. 15.

Larson, O. (2001). Charles Reep: A Major Influence on British and Norwegian Football. *Soccer & Society*, *2*(3), 58–78. https://doi.org/10.1080/714004854

Luzum, N., & Model, M. (2022). *Metrics – The Soccer Analytics Revolution*. https://sites.duke.edu/socceranalyticsrevolution/metrics/

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*(3), 109–118. https://doi.org/10.1111/j.1467-9574.1982.tb00782.x

Martinez Arastey, G. (2018). *Opta Sports: The leading sports data provider*. Sport Performance Analysis. Retrieved 26 January 2022, from https://www.sportperformanceanalysis.com/article/opta-leading-sport-data-provider

Mcparland, A., Ackery, A., & Detsky, A. S. (2020). Advanced analytics to improve performance: Can healthcare replicate the success of professional sports? BMJ Quality & Safety, 29(5), 405–408. https://doi.org/10.1136/bmjqs-2019-010415

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. *International Journal of Data Science and Analytics*, *5*(4), 213–222. https://doi.org/10.1007/s41060-017-0093-7

Morrison, J., & Rad, N. F. (2018) A Machine Learning Approach to Predicting Regular Season Success in the National Hockey League.

Moroney M J (1956). Facts from figures. 3rd edition, Penguin, London.
Murphy, A. H. (1969). On the" ranked probability score". *J. Appl. Meteor.*, *8*, 988-989.

Opta Event Definitions. (n.d.). Stats Perform. Retrieved 2 May 2021, from https://www.statsperform.com/opta-event-definitions/

Oshiro, T., Perez, P., & Baranauskas, J. (2012). How Many Trees in a Random Forest? In *Lecture notes in computer science* (Vol. 7376). https://doi.org/10.1007/978-3-642-31537-4_13

Partida, A., Martinez, A., Durrer, C., Gutierrez, O., & Posta, F. (2021). Modeling of Football Match Outcomes with Expected Goals Statistic. *Journal of Student Research*, *10*(1). https://doi.org/10.47611/jsr.v10i1.1116

Pollard, R. (2002). Charles Reep (1904-2002): Pioneer of notational and performance analysis in football. *Journal of Sports Sciences*, *20*(10), 853–855. https://doi.org/10.1080/026404102320675684

Prasetio, D., & Harlili, Dra. (2016). Predicting football match results with logistic regression. *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 1–5. https://doi.org/10.1109/ICAICTA.2016.7803111

Pratley, R. (2020). How Data Analysis won FC Midtjylland a title (and more). *Breaking The Lines*. https://breakingthelines.com/data-analysis/how-data-analysis-won-fc-midtjylland-a-title-and-more/

Pugsee, P., & Pattawong, P. (2019, August). Football Match Result Prediction Using the Random Forest Classifier. In *Proceedings of the 2nd International Conference on Big Data Technologies* (pp. 154-158).

Richter, F. (2021, February 5). Infographic: Super Bowl Pales in Comparison to the Biggest Game in Soccer. Statista Infographics. https://www.statista.com/chart/16875/super-bowl-viewership-vs-world-cup-final/

Robberechts, P., Van Haaren, J., & Davis, J. (2019). Who will win it? An in-game win probability model for football. *arXiv preprint arXiv:1906.05029*.

Rudrapal, D., Boro, S., Srivastava, J., & Singh, S. (2020). *A Deep Learning Approach to Predict Football Match Result* (pp. 93–99). https://doi.org/10.1007/978-981-13-8676-3_9

Saxena, S. (2021). Softmax | What is Softmax Activation Function | Introduction to Softmax. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/introduction-to-softmax-for-neural-network/

Schoenfeld, B. (2019). How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory. *The New York Times*. http://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html

Shmueli, B. (2020, July 3). *Multi-Class Metrics Made Simple, Part II: The F1-score*. Medium. https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

Shtovba, T. D., Tsakonas, A., Dounias, G., Shtovba, S., & Vivdyuk, V. (2002). Soft Computing-Based Result Prediction of Football Games. Proceedings of the First International Conference on Inductive Modeling, Lviv, 15–21.

Starkweather, J., & Moske, A. K. (2011). Multinomial logistic regression.

Trevisan, V. (2022). *Multiclass classification evaluation with ROC Curves and ROC AUC*. Medium. https://towardsdatascience.com/multiclass-classification-evaluation-with-roc-curves-and-roc-auc-294fd4617e3a

UEFA.com. (n.d.). *Country coefficients | UEFA Coefficients*. UEFA.Com. Retrieved 19 March 2022, from https://www.uefa.com/nationalassociations/uefarankings/country/

Ulmer, B., & Fernandez, M. (2013). *Predicting Soccer Match Results in the English Premier League*. 5.

UKEssays. (November 2018). The Changes Brought to Soccer by Big Data Analytics . Retrieved from https://www.ukessays.com/essays/technology/big-data-in-the-sport-of-soccer.php?vref=1

Wang, J., & Fan, Q. (2021, March). Application of Machine Learning on NBA Data Sets. In *Journal of Physics: Conference Series* (Vol. 1802, No. 3, p. 032036). IOP Publishing.

Weissbock, J., Viktor, H., & Inkpen, D. (2013, September). Use of Performance Metrics to Forecast Success in the National Hockey League. In *MLSA@ PKDD/ECML* (pp. 39-48).

Wheatcroft, E. (2020). A profitable model for predicting the over/under market in football. International Journal of Forecasting, 36(3), 916–932. https://doi.org/10.1016/j.ijforecast.2019.11.001

Wheatcroft, E., & Sienkiewicz, E. (2021). A Probabilistic Model for Predicting Shot Success in Football. *arXiv preprint arXiv:2101.02104*.

XG Explained. (n.d.). FBref.Com. Retrieved 30 January 2022, from https://fbref.com/en/expected-goals-model-explained/

Yusof, M. M., Fauzee, M. S. O., & Latif, R. A. (2014). Forecasting a winner for Malaysian Cup 2013 using soccer simulation model. 1153–1159. https://doi.org/10.1063/1.488775