Aalto University
School of Business

How Accurately Does the Expected Goals Model Reflect Goalscoring and Success in Football?

Bachelor's Thesis
Tuomas Tiippana
xx.xx.2020
ISM Program

Approved in the Department of Information and Service Management
xx.xx.20xx and awarded the grade

Author Tuomas Tiippana

Title of thesis How Accurately Does the Expected Goals Model Reflect Goalscoring and Success in Football?

Degree Bachelor's degree

Degree programme Information and Service Management

Thesis advisor(s) Advisors

Year of approval 2020 **Number of pages** 28 **Language** English

Abstract

This thesis discusses the expected goal model for football and assesses the explanatory power of the model to estimate match results and score lines as well as success at the end of the season. The objective of this thesis is to evaluate the expected goals model in order to fill a gap in academic literature. By filling this gap, the model and its outputs can be more accurately used in the media as well as in real world problems where this model is applicable. Additionally, in this thesis attempt is made to improve the results the expected goals model produces by running the estimated scores through the Poisson distribution. This is done since the Poisson distribution is frequently used in the academic literature on the field as it has been observed that goals in single matches in football are Poisson distributed.

In addition to the Poisson distributed expected goals model, two other models estimating the outcomes and score lines of individual matches and two probability-based methods were used as benchmarks. Shots on target-based model and naïve aggregate model were chosen as simple models to estimate the results and scores of individual matches. Market odds probabilities and probabilities produced by the Elo ranking system were used to further benchmark the performance of the expected goals model. The data used in this study is collected from the four biggest leagues in football: The English Premier League, Spanish La Liga, German Bundesliga and Italian Serie A. The seasons included in the study were the five seasons from 2014-15 to 2018-19, totalling 7230 matches. The data were collected into and modified with Excel.

Comparisons indicated that while none of the models could predict or estimate the results and score lines of single matches particularly well, the Poisson distributed expected goals model came the closest with a minimal difference to the standard expected goals model. Additionally, it was found that the model is biased in estimating an excessive number of draws, while underestimating ,the amount of matches that end with either or both of the teams failing to score. However, when it came to estimating the success of teams, i.e. the amount of accumulated points by the end of the season, the standard version of expected goals fared the best with its derivative output expected points. All in all, the results imply that while it is extremely difficult to accurately model football match outcomes, the expected goals model can give some insight behind the score line and is a valuable analysis tool.

Keywords Expected goals, football analytics, sports analytics

Table of Contents

1 Introduction

Analytics, statistics and the use of data in many different forms have taken the sporting world by storm in recent decades and has arguably been one of the most significant revolutions within the industry across the globe (Severini, Thomas 2012). New and improved technologies can quantify actions from the field of play into numerical form, giving rise to sports analytics. Association football, however, has not been leading the data revolution in sports. Despite being clearly the biggest sport in the world, football has not been able to match the level of quality of analytical modelling of other majority sports (Brooks, Kerr, and Guttag 2016). On the contrary, recent results and emotions are still the biggest factors in decision making in football rather than data, which is the case in other professional sports (Eggels 2016).

Football is a game of high complexity, resulting from its low-scoring and dynamic nature. This makes it extremely difficult to model. Today the most common statistics used to assess team performance and match events are very rudimentary, such as shots, shots on goal, ball possession and the number of passes (Brooks, Kerr, and Guttag 2016). These statistics fail to portray an accurate image of a game and explain the outcome as they offer too constructed view on the game. A game as complex as football demands both quantitative and qualitative performance indicators.

An innovative solution to this problem of modelling the game realistically was presented by Sam Green (2012) in his article "Assessing the performance of Premier League goalscorers". Green developed an analytical model called expected goals (also referred to as xG) based on goalscoring probabilities from certain goalscoring opportunities. The idea of this model was to assign a probability of a goalscoring chance being converted into a goal. For example, if a goal scoring opportunity is assigned an xG value of 0,3, the effort is converted into a goal 30% of the time with average finishing. By adding up all the probabilities from the chances created during a football match we can get an outcome that describes how the match should have ended with average finishing.

1.1 Introduction to the expected goals model

The expected goals model is a neural network prediction algorithm that uses over 10 parameters to produce the output (Understat.com, 2020). These include for example distance and angle to the goal, the body part used to strike the ball, situation (open play,

penalty, freekick, corner etc.), possible errors committed in the scoring situation, the number of dribbled players before the effort, and the state of the game, i.e. is the game even or is the home or away team leading. The model has been trained using over 100 000 efforts on goals.

The model has three main use cases in professional football. Firstly, to assess the performances of the teams playing. Since football is a low-scoring game, luck often plays a role in the eventual outcome. As a result teams that played very well, created multiple chances to score and were by all possible standards the "better team", may end up losing. The expected goals model gives us insight to the game beyond the score line: Which team was more likely to win the game in the light of created goal scoring opportunities, and second of all, how well could the teams convert these opportunities into actual goals. In the table 1.1 the observed and expected goals of Manchester City and Liverpool are displayed. Even though Manchester City failed to create more goal expectancy than Liverpool, they managed to convert these chances into goals and eventually won the match. In light of the created goal expectancies, Manchester City could be considered lucky to have won the game. In figure 1.1 all the teams playing in the Premier League during the 18/19 season have been plotted on a graph to visualize how well they have managed score goals in retrospect to their xG figures. The linear equation represents scoring the exact same amount of goals that the team's xG value suggests. Manchester City and Liverpool can be seen to have created by far the biggest xG values. Liverpool has been the most efficient when it comes to finishing, while Fulham has had troubles converting chances into goals.

Table 1.1: The observed and expected goals of a match between Manchester City and Liverpool on 3rd of January 2019.

	Manchester City	Liverpool
Actual score	2	1
Expected goals	1,00	1,38

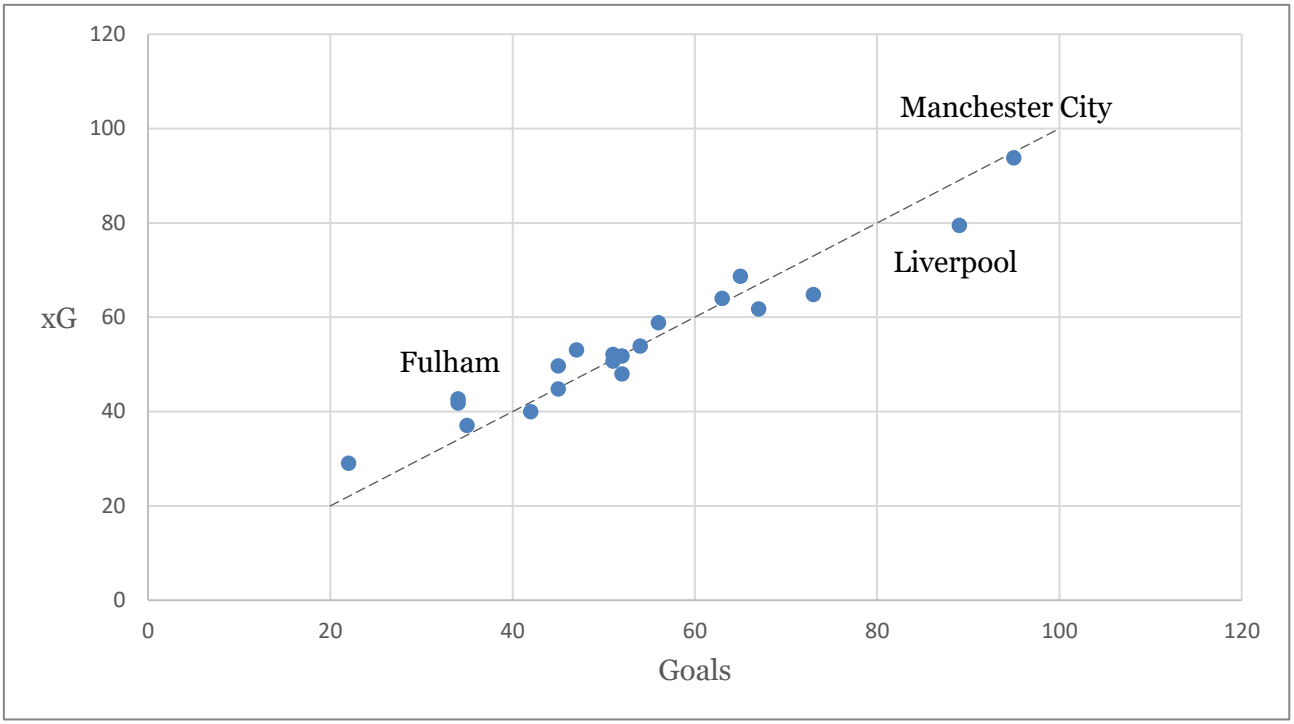

Figure 1.1: The observed and expected goals for every Premier League team from the 18/19 season.

The xG values can be examined from the defensive aspect of the game as well. If opponents can generate sizeable expected goals against values against a certain team, the management of the team in question can conclude that they have to improve their defensive quality and prevent opponents into getting dangerous goals scoring situations. In addition, if a team concedes more goals than its expected goals against figures suggest, their defensive players might be underperforming in critical situations.

The second use case of the expected goals model is to analyse the performance of individual players. Each player gets their own expected goals value and expected assists value. From these figures we can extract three objects of player performance analysis. Firstly, how much did a certain player add to his team's odds of scoring a goal during the game, which can be seen straight from the player specific expected goals value. Secondly, how effectively could the player convert these probabilities into a goal. This can be calculated by detracting the amount of goals scored from the goal expectancy of the player. Finally, the model indicates how well could the player assist his or her teammates' into getting better goal scoring opportunities. This is expressed as the expected assists value which is a derivative of the expected goals model.

The third and final use case of the model is probably the most relevant in terms of actions that can be taken based on it: Recognizing possible transfer targets to recruit in order to strengthen the team. Many clubs have huge analytics departments and scouts all over the

world keeping tabs on exiting players who could be acquired to bring either goal scoring efficiency or the ability to create goalscoring opportunities for them. For these purposes analysing the goal expectancy, efficiency of converting chances into goals and expected assists valuable information about the player can be gained in a unbiased and reliable way.

1.2 Research objectives and research questions

The objective of this research is to assess the ability of the expected goals model to reflect football as a game: How well does it explain the score line and can coaches and analysts trust the results it produces. The model has had a lot of media attention, created arguments both for and against it, and there is still plenty of uncertainty on how to interpret it and whether or not to trust its results. In this thesis, the expected goals model is explained, analysed and evaluated in order to offer some clarity to the details and accuracy of the model and to create scientific grounds for the use and interpretation of the expected goals model. Five other models are also included in the research to function as benchmarks to further examine the model's performance.

Additionally, the expected goals model is fitted with Poisson distribution that will be benchmarked against the original model to see whether the explanatory power can be enhanced.

The main research questions which are covered in this thesis are:

1) How accurately the expected goals model is able to explain match outcomes and success over a season?
2) Can the accuracy of the model be improved with a Poisson distribution?

1.3 Scope of research

The scope of this thesis is limited to assessing the explanatory power of the expected goals model. A comprehensive analysis of how accurately the model describes and reflects the events and nature of the game is composed. As the objective of this paper is to determine to what extent this new analytical model is capable of representing the actual balance of power between of two teams playing a game of football, this paper will go in detail to analyse the outputs of the model. Altogether, the aim of this thesis is to

provide the reader with a comprehensive understanding of the expected goals model and give tools to critically evaluate and interpret the results it produces.

This thesis will neither cover the statistical and mathematical methods used to construct the model nor will it try to build a similar model like the majority of scientific literature on the subject has done (Eggels 2016; Macdonald 2012).

1.4 Structure of the research

This thesis is divided in five chapters. In chapter 2, previous literature is reviewed and theoretical framework for the use of Poisson distribution in football and goodness-of-fit and error measurements is provided. The data sources, benchmarking methods and model evaluation methods are justified in chapter 3. Chapter 4 presents the results of the study. In chapter 5 the results are further discussed, and the thesis concluded. Lastly, the limitations of the study are discussed, and future research topics are suggested.

2 Theoretical background

In this chapter the theoretical background for this thesis and its topics are presented. In part 2.1, previous literature on expected goals is introduced and discussed. In part 2.2 the reader is presented with the idea of fitting a Poisson distribution to the outputs of the model to improve its accuracy. The probability distribution itself is explained and its application in football context is supported with previous literature. Finally, in part 2.3 the measurements of error used to compare models and benchmarks used in this thesis are presented and their validity is established through both the academic literature on the subject and the factors in this particular research.

2.1 Academic literature on the expected goals model

In this chapter the previous literature as well as the recent development of the model and its derivatives are discussed. In academic literature, predicting and explaining the result of a football match has been a major topic of interest. The interests for predicting match outcomes arise mostly out of betting and attempting to beat the market odds. Consequently, a major share of the academic literature on match forecasting approaches the subject from betting perspective (Cain, David, and Peel 2000; Constantinou, Fenton, and Neil 2013; Nielsen and Sandøy 2019). There are, however, a few articles related to the expected goals model, even though a large proportion of research on the subject is conducted privately by companies focused on sport analytics and data, such as Opta Sports, or football clubs pursuing financial and sporting success. Outside of football, a study in ice hockey by Brian Macdonald (2012) has been the only effort in adapting this kind of assumed probabilities of goalscoring in the field of sports.

Like stated in the chapter 1.1, Sam Green came up with the initial idea to assign a probability to a certain goal scoring attempt in 2012. Since then his work has been continued and extended mainly by statisticians in their personal blogs (Rathke 2017). In addition to these blogs, there has been apparently only a few academic studies on the subject of the expected goals (Eggels 2016; Lucey et al. 2014; Rathke 2017) of which all have tried to construct a model that explains individual match outcomes, rather than actually seeking validation for such models through benchmarking and comprehensive evaluation. For example, in Eggel's study from 2016, the only metrics reported from the model he built were the correct result percentage and the correct score line percentage the model achieved. In the study conducted by Rathke, only the coefficient of

determination between the expected and observed goals of the English Premier League and German Bundesliga teams and players was reported. These two examples demonstrate well how the field lacks comprehensive analysis of the reliability and validity of the model.

Lately, the literature around the subject of expected goals has extended to asses the values of passes (Power et al. 2017) and the entire attacking phase even if it does not result in a goal scoring attempt (Fernández, Bornn, and Cervone 2019; Link, Lang, and Seidenschwarz 2016; Spearman 2018). In all of these models the expected goal value is of essence and the entire models are more or less built around the eventual goalscoring probabilities.

2.2 The use of Poisson distribution in football

In football context the goals scored by each team in a single match are often considered to follow a Poisson distribution. A plethora of studies have used this approach in modelling and predicting goals in football (A. Heuer, Müller, and Rubner 2010; Jones, James, and Mellalieu 2004; Karlis and Ntzoufras 2003). The Poisson distribution demands that the events studied are discrete and they occur randomly, the mean of the events is known and finally, that one event is independent from another. In this context we study goals in individual matches which are, of course, discrete as you cannot score for example one third of a goal. On average in the leagues and seasons used for data in this study, a team scored 1,38 goals per game, while the interval of scoring a goal is naturally random. Finally, scoring a goal does not affect the probability or the time of scoring the next goal. However, those who follow football might argue that scoring a goal in a match may "stir the pot" and result in more goals either for the team that just scored or for the opposing team, but this has not been statistically proven. Based on this reasoning, the Poisson distribution was deemed to be acceptable when modelling the number of goals scored in a single match.

As the previous studies indicate and as the general consensus on the topic is that goals in football follow, in fact, the Poisson distribution, an additional model referred to as the xG Poisson model is created. In this model the outputs of the standard expected goals model are ran through the Poisson distribution. As a result, the most probable amount of goals for each team according to the distribution is derived and used as an additional model in this study.

2.3 Error measurement and model evaluation

As the eventual goal of this study is to determine how accurate and reliable the expected goals model is when it comes to reflecting goalscoring and success both in individual matches as well as during the entire season, it is of essence to be precise when choosing the methods to estimate this. There is a great deal of different error and correlation measurement methods that compare the predictive and explanatory power of models and benchmarks. In the following chapters these methods are discussed, and their attributes are evaluated in order to determine the best possible selection of comparison methods for this study. However, since neither the Elo rating model nor the market odds produce predictions for the score line but only for the result of individual matches, the error measurements cannot be applied to these benchmarks in individual games. As the comparisons of models for individual matches go, only the quantities of correct results can be analysed for these two models.

One of the most widely used statistic to assess the goodness-of-fit of a model is the coefficient of determination, or R^2 (Krueger and Lewis-Beck 2006). It gives a number between 0% and 100% that explains the percentage of total variation the model explains. While it has been somewhat of an industry standard, there has been debate on the subject whether the R^2 is a meaningful goodness-of-fit statistic or not. There has been arguments made by researchers that it does not add any value while assessing the validity or reliability of a model (King 1990a, 196; Aachen 1990). On the other hand, it is very intuitive with its interpretation and ultimately represents well the correlation between two sets of data.

In order to further evaluate and compare the models, error measurements are applied to the study as well. There is no optimal error measurement for every need, and that is why when selecting error measurements the features of the data one is dealing with sets the guidelines for error measurement selection. One of the major selection criteria is the outlier protection. Error measurements that are based on the sum of squared errors are inflated if the data set contains outliers or otherwise large values (Armstrong and Collopy 1992; Willmott and Matsuura 2005). The root-mean-squared-error (RMSE) and mean-squared-error (MSE) are the most popular examples of this kind of error measurements. Applying error measurements such as the RMSE to data sets containing outliers would produce biased outcomes and thus, could result in incorrect conclusions about the preferred model.

Another criterion for the methods are the zero-values found in the data. In the case that there are zero-values in the observed data, the methods based on absolute-percentage-errors (APE) cannot be applied, since in their formula the absolute difference between the observed value and expected value is divided by the observed value. In football, matches often end with one or both of the teams failing to score and since this would result in an observed value of zero, these methods cannot be used. However, when the models are compared by their performance of estimating the eventual point accumulation the problem of zero-values is no longer valid. Armstrong and Collopy (1992) recommend the use of MdAPE when choosing a method to compare forecasting models with large sets of data. They justify their recommendation with reliability, construct validity, informativeness and outlier protection. One negative feature with the APE based methods, however, is their asymmetricity, i.e. proneness to be biased towards low forecasts (Armstrong and Collopy 1992; Chen, Twycross, and Garibaldi 2017).

Willmott and Matsuura (2005) among others argue in favour of using the mean-absolute-error (MAE) as a natural measure of an average error. Like the name suggests, it is a very simple method of comparing forecast models. In all of its simplicity lies its power: It offers satisfactory protection against outliers and large errors, is easy to understand and interpret, and it is not biased.

3 Methodology

In this chapter the methodological framework used to analyse the expected goals model and the benchmarking methods are introduced and discussed. The chapter is divided in three parts. In the first part the validity and reliability of the data as well as the data collection methods used in the thesis are presented. The second part introduces and justifies the benchmarking methods used in this research. In the third and last part of this chapter the choices of measurement used to compare and analyse the expected goals model as well as the benchmarking models are explained and validated.

3.1 Data source and modification

The statistical models described and analysed in this paper demanded acquisition of data from multiple sources. The expected goals data used in this thesis is from Understat (Understat.com, 2020) which is a web site focusing on collecting, calculating and visualizing expected goals data from six national leagues, of which four were chosen for this thesis: English Premier League, Italian Serie A, German Bundesliga and Spanish La Liga. Understat was deemed to be a reliable data source since multiple academic papers and studies were using its data as their source for expected goals data (Flepp and Franck 2019; Andreas Heuer 2020; Nielsen and Sandøy 2019; Robberechts 2019). Because Understat does not provide data from seasons before the 2014-15 season, the seasons 2014-15 to 2018-19 from the aforementioned leagues were analysed in this paper. In total, this equals to 20 seasons and 7230 individual matches. Because Understat does not offer the possibility to download their data directly, a web crawler was used to collect the data into Excel. In addition to the expected goals data gathered from Understat.com, individual match data as well as the league tables were also acquired from Understat.

Besides Understat, two other sources for data were used. The data collected from the following web sites were used for benchmarking purposes. The shots on goal data that was acquired for the purpose of creating a benchmarking model for the expected goals model as well as the betting odds for individual matches were provided by Football-data.co.uk, a football betting site that offers historical data from 36 football leagues from multiple seasons. The betting odds for these individual matches are from up to 10 market leading bookmakers and the odds used in the thesis are the average odds offered by these companies. The data on Football-data.co.uk is downloadable in CSV format and thus did not require additional effort in modifying it.

Finally, the Elo probabilities for each match were obtained from Elofootball.com, a football statics website dedicated on collecting and portraying Elo rankings and probabilities for individual matches derived from these rankings. The data from this site was obtained through a web crawler.

3.2 Benchmarking methods

In order to analyse the credibility and validity of the expected goals model, benchmarking was used to compare various different models. In total, there were four models or methods used in this thesis to predict and explain match outcomes and success over a season. These four methods were the shots on target model, naïve aggregate model, the Elo rating model and average betting odds for each match. The benchmarking methods were chosen to represent the vast spectrum of data and methods used nowadays to model match outcomes and success.

The first benchmarking model is the shots on target. The model was chosen because it captures the goal scoring efforts from a quantitative aspect, in comparison to the expected goals model that aims to assign a qualitative value as well for the scoring opportunities. The model is ultimately a very simple one. It multiplies the number of shots on target the team managed to produce in the game with the number of goals scored throughout the season divided by the shots on target over the season. This is illustrated in the following formula:

$$Goals\ scored = Shots\ on\ target * \frac{Goals\ scored\ during\ the\ season}{Shots\ on\ target\ \ throughout\ the\ season}$$

The second method chosen was the aggregate model. The model calculated both home and away attacking and defensive strengths for each team based on scored and allowed goals home and away from the season in question. The optimal way of building the model would have been by using the data of the previous season, but that approach was deemed unfair because the statistics for the teams that were promoted for the current season would have been massively too optimistic. A compromise was made and as a result the model should reflect well the balance of power between the teams in individual matches and in the eventual accumulated points.

The third benchmark method applied is the Elo rating model (Elo, 1987). The reasoning behind choosing the Elo model as a benchmark comes from its stature in the sport analytics world. It is one of the most widely adapted rating and prediction models and

while it was originally meant for rating chess players, it has since been applied to multiple other sports as well, such as ice hockey, basketball and football (FIFA 2018; Štrumbelj and Vračar 2012; Tenkanen 2019). Even the most famous ranking system in football, the FIFA World Ranking table, which ranks all the national teams in the world, has adopted the Elo rating system. (FIFA 2018). It is an adjustive rating model that continuously updates the ratings of all teams based on their results, emphasizing the latest results. The ratings of each individual teams are then converted into probabilities for each possible match result, i.e. home team win, draw, away team win. However, as the Elo model can only predict the match result, the score line predictions for individual matches cannot be made.

The fourth and final benchmark used is the market betting odds. In previous academic papers the market odds have been proven to be a superior way of predicting match results (Hvattum and Arntzen 2010). Because of its proven track record of predicting match results, the market odds was included as a benchmark method in this thesis. Similarly to the Elo rating system and its predictions, the market odds do not produce score predictions.

In addition to these four methods, the xG Poisson model was also applied to this thesis. This model was built by taking the estimated score produced by the expected goals model and calculating the probability of each team scoring a certain number of goals through Poisson distribution. As a result the most probable result of the match was derived.

To better illustrate the outcomes produced by each model, the following table shows the estimated results and scores for a match between Arsenal and Crystal Palace played on the 21st of April in 2019. The match resulted in a 2-3 home defeat for Arsenal.

Table 3.1: The estimates and prediction of the models used in this thesis for the game between Arsenal and Crystal Palace on 21st of April 2019.

	Arsenal	-	Crystal Palace	Correct result	Correct score
xG	1,54		2,56	Yes	Yes
xG Poisson	1		2	Yes	No
Shots on target	2,05		2,44	No	No
Aggregate	2,23		1,13	No	No
Elo	83 %		17 %	No	No
Odds	66 %	22 %	16 %	No	No

Each of the models used for benchmarking produced an outcome for each match. Based on these outcomes, points were divided between the teams according to the standard method in football: 3 points for a win, 1 for a draw and 0 for a defeat. Moreover, for the expected goals model the accumulated points were calculated with two slightly different methods, both derived from the same outputs. The model has an output called expected points, which divides the points between the two teams on a continuous scale from 0 to 3. In the match between Arsenal and Crystal Palace the expected points were divided 0,71 - 2,09, representing the goal expectancies and probabilities of winning the match. However, for benchmarking purposes the accumulated points for each team were decided to calculate by granting the points discretely between the teams as well, like described above.

One compromise with the benchmarking methods is that while the Elo and market odds methods are purely predictive, the rest of the models are not. However, like stated before, football is extremely difficult to model since it is such a complex game and thus, these different benchmark methods were deemed to be acceptable. The different approaches the methods present on predicting and estimating the result and score line can be seen as a positive issue as well since they demonstrate which factors affect the result the most: The events on the field, the quality of the teams or a mixture of these. The market odds and Elo ratings base their predictions on the strengths of the teams. Additionally, the market odds method gives a prediction that includes all the information available about the event: The fortes of both teams, injured and suspended players, the context of the game, are the teams fighting for survival or the championship or just merely playing without any particular stake. The expected goals and shots on target models both estimate the result entirely based on what has happened on the field while the aggregate model evaluates the strengths of the teams based on their success during that exact season.

3.3 The chosen methods for model evaluation

In order to effectively assess how accurately the expected goals model and the benchmarks reflect the game and explain the results and score lines, there is a number of error and correlation measurements that can be used. However, each of these methods have their pitfalls as well as optimal use cases and thus, the purpose of this chapter is to determine the ones that are most suitable for the needs of this particular study. The error measurements were chosen to give as broad aspect as possible to reflect the predictive and explanatory power of the models. Like stated in the chapter 2.3, there are two

instances where the models are compared to each other: The explanatory power of the models when estimating goalscoring in individual matches, and when modelling the eventual accumulated points at the end of the season. Since the two instances where the measurements are applied to are vastly different, the error measurements chosen for them naturally differ as well. In the following chapters the reasoning for the selection of each error measurement is presented.

The first method chosen to measure if the predicted outcomes of the models' are, in fact, in line with the actual observations was the coefficient of determination, or R^2. It gives a straight-forward and informative number that represents the linear correlation between the observed and expected points of data, which indicates how well the model can explain the total variation in the observed data set. On the other hand, like stated in the chapter 2.3, the coefficient of determination does have its critics. Legates and McCabe (1999) argue that because high correlations can be achieved by any model, good or poor, correlation-based measures like the coefficient of determination should not be used when evaluating models. Despite of this comment, the R^2 was deemed to give valuable information about the validity of the models and was chosen as one of evaluation methods for both phenomena studied in this thesis; The goals scored during single matches and points accumulated at the end of the season.

When choosing the methods of measuring the error terms of observed and estimated values a number of factors were considered. First of all, the determinant factor when selecting error measurement methods is the quantity and quality of outliers, or more particularly large values in the data that might result in biased evaluation results. This was done since some measurement methods, such as the root-mean-square-error, do not work well with data that has exceptionally large values or outliers (Legates and McCabe 1999; Willmott and Matsuura 2005). The reasoning to react cautiously to these large values is that often in football luck or individual errors might result in an disproportioned number of goals. For this purpose, an outlier analysis with box plots was conducted.

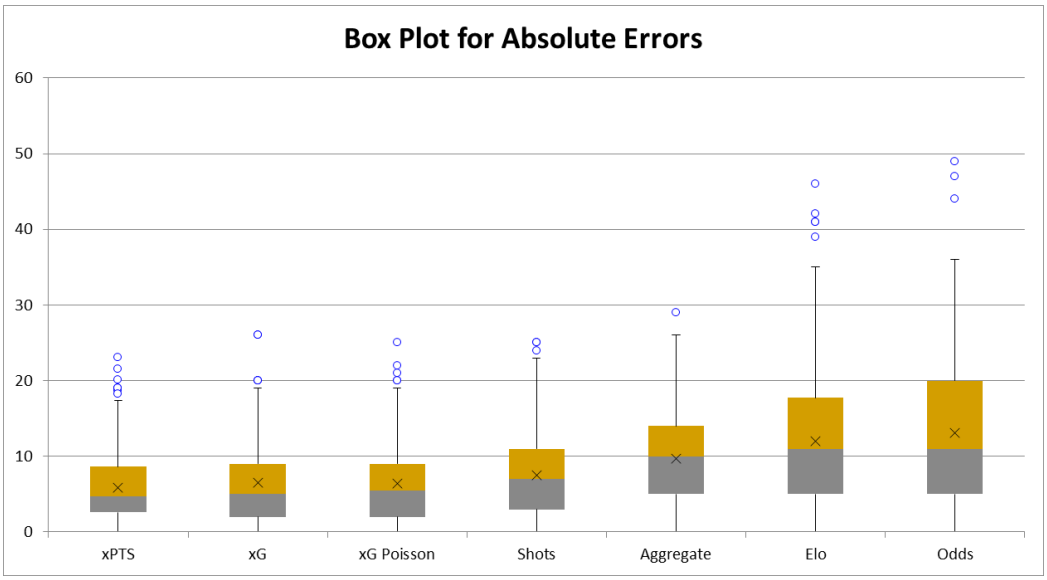

Figure 3.1: Box plots for absolute errors of cumulated points for the models and benchmarks.

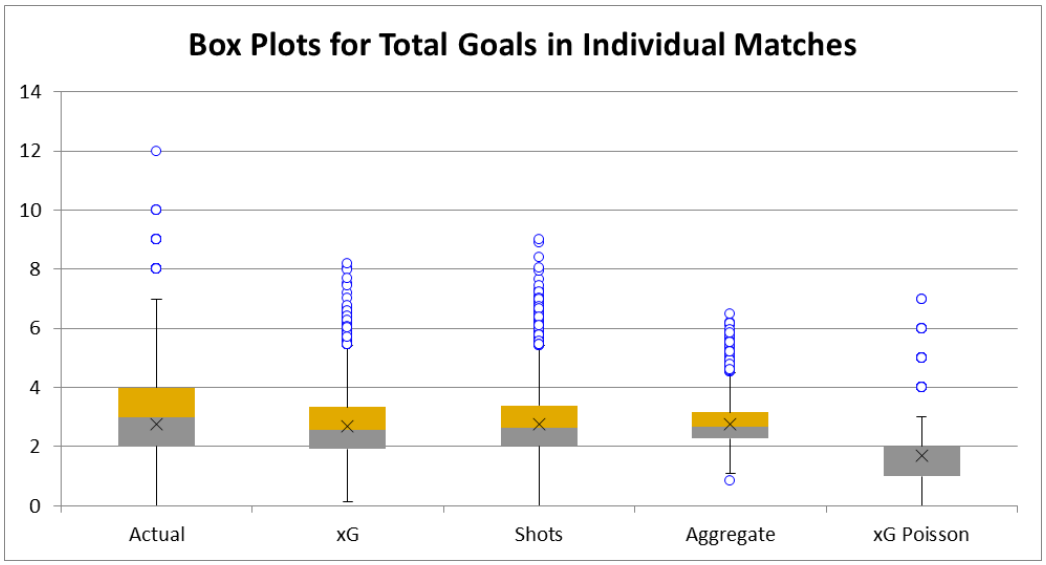

Figure 3.2: Box plots of total number of goals scored in individual matches.

As can be seen from the graphs above, both sets of data contain these unusually large values. Noteworthy notion from the figure 3.2. is the number of outliers the xG Poisson model has (460). Since all the outliers are on a single discrete value, such as 4, 5, 6 and 7, the plot does not visualize the number of outliers well. Since RMSE as well as other methods based on the sum of squared errors cannot handle outliers properly, they are excluded from the selection of error measurements.

The second method of measuring model performance chosen for both evaluation of goal scoring in individual matches as well as the eventual success in accumulated points is the mean-absolute-error, or MAE. The rationale for this is on one hand its simplicity and un-

biasedness, and on the other hand its resistance from outliers (Legates and McCabe 1999; Willmott and Matsuura 2005). In the context of scoring goals and accumulating points, the mean-absolute-error produces highly informative values. For example, the MAE of expected goals method when estimating the number of goals scored in a single match is 1,4 which gives a clear idea of both scale and amplitude of the error the model produces.

The third and final error measurement applied to forecasting the accumulated points at the end of the season is the median-absolute-percentage-error, MdAPE. Armstrong and Collopy (1992) suggested MdAPE for forecast method comparisons for its reliability, validity, resistance against outliers and informativeness. It was preferred over its close relative, MAPE, because the median-based model handles outliers better than mean-based.

In summary, two model evaluation methods are used to compare the estimates and predictions of number of goals scored in individual games: MAE and the coefficient of determination. When evaluating the same models but from the aspect of accumulated points at the end of the season, a third method, MdAPE, is added. Moreover, observed and modelled means as well as standard deviations are presented for further information on the descriptive statistics for the reader, as suggested by Legates and McCabe (1999).

4 Results

In this chapter the results of this research are presented. This chapter is divided in three parts. In the first part the results and score lines the different models produce are evaluated to give some insight into the performance of different models. The second part continues the assessment of model performance in individual games by presenting the results of the explanatory power of the models in individual matches through analysing the results of model evaluation methods. As stated in the chapter 3.2, not all models are capable of producing score line estimates and thus, the Elo model and market odds cannot be compared as comprehensively as the others.

In the third part the results of how different models fare in estimating the eventual accumulated points, i.e. the eventual success of the teams, are presented. Before assessing the results of the coefficient of determination and error measurements, the residuals are analysed in order to further analyse if the models used are biased or not. Majority of conclusions are discussed in chapter 5.

4.1 Match outcomes

Perhaps the most intuitive approach to assess the reliability and validity is to evaluate the outputs of the models compared to the observed results and scores. In table 4.1 the percentages of correct outcomes and scores produced by the models as well as correct scores that were one goal off are presented.

Table 4.1: Match outcome evaluation

Model	Correct outcome %	Correct score %	Correct score +/-1 %
xG	57 %	17 %	57 %
xG Poisson	59 %	18 %	53 %
Shots	47 %	11 %	44 %
Aggregate	54 %	13 %	48 %
Elo	53 %	-	-
Odds	54 %	-	-

With the Poisson distributed expected goals model, both the correct outcome and correct score percentages improved in comparison to the standard expected goals model. As stated in chapter 3.2, the values of the aggregate model are a bit unfair, since it gets its data from the season at hand. Traditionally the market odds have been the best predictor

of results, and this notion cannot be nullified since excluding the Elo model, all other models estimate the result and score based on the events of the match (Constantinou, Fenton, and Neil 2013; Hvattum and Arntzen 2010). Thus, while market odds and the Elo model aim to predict the result, other models are trying to explain it.

The modest correct score percentages the models produce are expected. As football is by its nature a low-scoring game, both the results and scores can be affected by a stroke of luck in individual games. At the end of a season, these coincidences and strokes of luck should even out and models such as the expected goals model should produce more accurate results. Lastly and rather surprisingly, the expected goals model performs the best when estimating the correct score when given a tolerance of one goal off in the estimation. Intuitively the xG Poisson model should have a higher percentage since it fares better in the other two measures. However, by applying the Poisson distribution to the score estimates often effectively rounds the values downwards which more often leads to the correct outcome but is off from the actual score by more than one goal. A good example would be the game between Aston Villa and Chelsea, played on the 7th of February 2015. This effect can also be seen in the box plots of figure 3.2.

Table 4.2: The xG and xG Poisson estimates of a match between Aston Villa and Chelsea (7.2.2015).

	Aston Villa	Chelsea
Actual Score	1	2
xG	0,71	1,12
xG Poisson	0	1

While the match ended in Chelsea's 2-1 away victory, the expected goals model estimated it to end in a 1-1 draw. However, the Poisson fitted expected goals model got the result correct but failed to estimate the score as accurately as the standard expected goals model.

Another way to evaluate the results of the expected goals model is to compare the score line estimates to the observed ones. In figure 4.1 a histogram for the 11 most common score lines from the date set are presented. In the histogram both the number of observed and expected score lines are visible. The expected goals model struggles to estimate the number of games that end with either of the teams failing to score. The most notable error is the difference between observed and expected games resulting in a 0-0 draw, 469 instances. Another peculiarity is the amount of 1-1 draws, the difference between the two being 975.

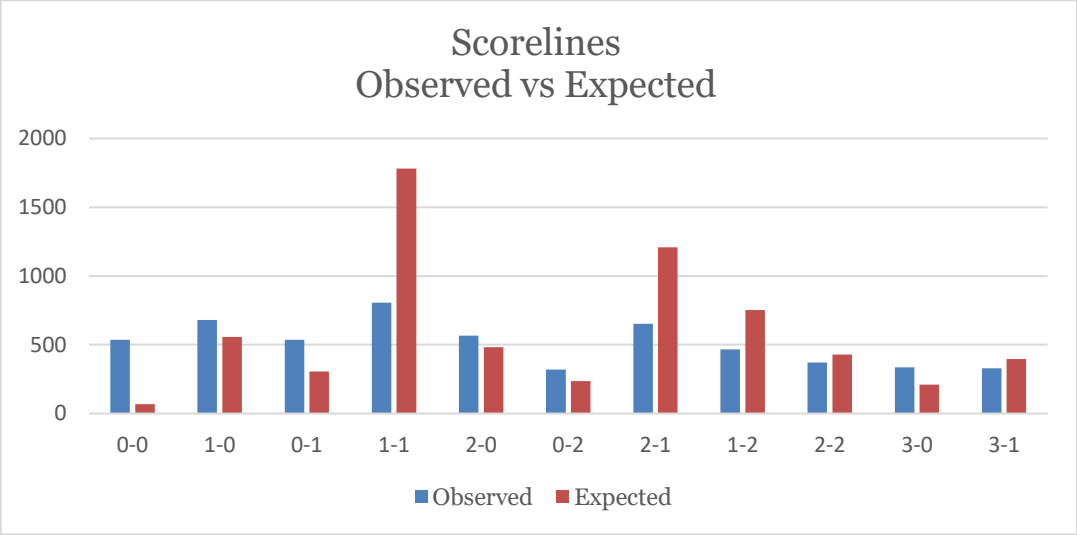

Figure 4.2: Histogram of the 11 most common observed and expected score lines according to the xG model.

One interesting phenomenon the score line analysis displayed was the amount home defeats in comparison to the actual observed home defeats. In table 4.3 the observed and expected results are shown. The home win percentages are roughly the same, but while in reality 25% of all games result in a draw, the expected goals method over emphasized this figure. Out of all the tight contests bound to end in a draw, in many cases ultimately the away team could score a goal and secure an away victory. This can be even further demonstrated when taking all the tight games, i.e. the games that ended with a goal difference of 1 or 0 and calculating the average error between the scored goals and expected goals for the home and away team. On average, the error term for the home team is -0,20, or -17% of the average number of goals scored by the home team, and -0,04, or -4% for the away team.

Table 4.3: The observed and expected results of all matches from home team's perspective.

Result	Observed	Expected
Win	45 %	44 %
Draw	25 %	32 %
Defeat	30 %	24 %

4.2 Results on explanatory power in individual matches

For each game of the four leagues across five seasons, six different results and four different score lines were produced. These score lines were then compared to see which method can model the observed outcomes most accurately. For individual matches, two metrics were used to evaluate the models, mean-absolute-error and the coefficient of determination, i.e. the R^2. Table 4.4 presents the results.

Table 4.4: The MAE and R^2 of the models and benchmarks.

Model	MAE	R^2
Expected Goals	1,49	40,2 %
xG Poisson	1,56	36,2 %
Aggregate	1,68	26,3 %
Shots on Target	1,87	12,9 %

The table is ranked according to the MAE value of each model to make comparing the figures easier. Even though the xG Poisson model was more accurate in estimating the results and score lines of individual games, the standard expected goals model performed better both with the MAE and R^2. A logical explanation one could come up with is the granularity of the expected goals method, but when the same error measurement and R^2 were applied to the rounded version of the model, it produced even better MAE (1,42) yet worse R^2 (35,9%). Similarly how the standard expected goals model produced essentially more accurate estimates of score lines than other models, the comparisons made with MAE produced the same result. The R^2 value was as low as one could expect since, similarly to the *correct score percentage*, the effect luck has in individual games is considerable.

4.3 Results on explanatory power in estimating success

In this chapter the models and benchmarks are evaluated on how accurately they can model the ultimate success during one season, the accumulated points. The evaluation is done with three methods: MAE, MdAPE and R^2. As explained in chapter 3.1, a new output of the expected goals model called expected points (xPTS) is taken into the comparison as well. Before evaluating the models any further, a residual plot analysis is made in order to reveal possible patterns in the residuals. By assessing whether the models produce any specific residual patterns or not, we can determine if the models are

biased. The residual plots for the expected points, the expected goals model and the Elo model are displayed in the figure 4.2.

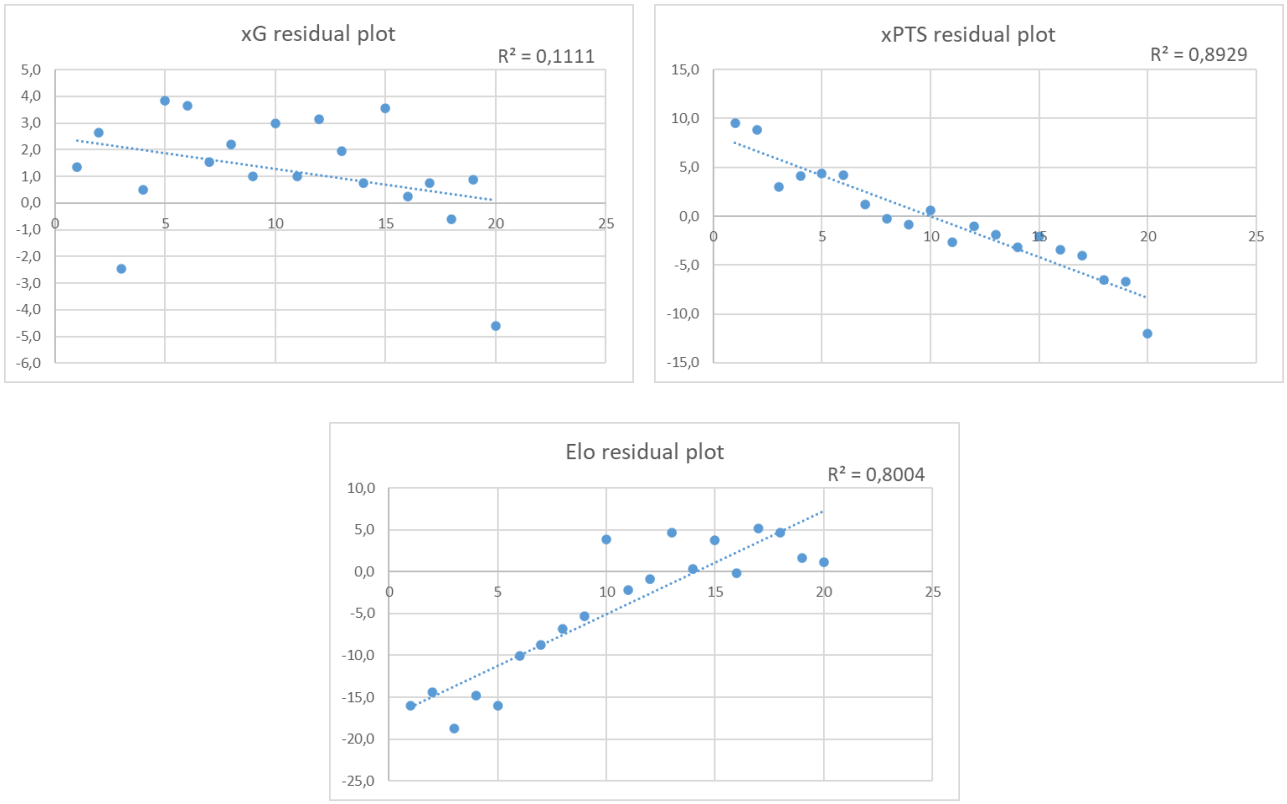

Figure 4.2: The residual plots for accumulated points of expected points, expected goals model and the Elo model. On the x-axis are the league table positions.

The three residual plots were chosen since they represent all three different kinds of residual plots the models produced. The expected points residuals form a distinct downward trend and the R^2 value is notably high. This indicates that top teams accumulate more points on average than they should have while bottom teams gathered less, according to the expected points method. On the other hand, both expected points and its xG Poisson counterpart did not create any notable patterns. The rest of the models form an upward trend of residuals, like can be seen in the residual plot of the Elo model. This indicates that according to these models, top teams should have earned more points than they did. Again, this can be seen to be caused by the low-scoring nature of the game that allows worse teams to beat better teams by a single stroke of luck or a piece of skill.

It is, however, critically important to know whether the models that are being evaluated are biased or not to be able to interpret them and apply them to real world situations. In table 4.5 the goodness-of-fit measure R^2, and the two error measurements MAE and

MdAPE are shown. The means and standard deviations are also shown in order to better demonstrate the performance of models.

Table 4.5: R^2, MAE, MdAPE, means and standard deviations of the models and benchmarks.

Model	R^2	MAE	MdAPE	Mean	Std deviation
xPTS	84,1 %	5,9	9,4 %	51,23	13,69
xG	81,3 %	6,5	11,1 %	49,73	18,65
xG Poisson	81,7 %	6,4	10,7 %	50,04	18,57
Shots on Target	80,9 %	7,6	14,1 %	48,93	20,41
Aggregate	91,7 %	9,7	18,2 %	48,71	26,73
Elo	79,0 %	12,0	19,6 %	55,62	27,46
Odds	78,9 %	13,1	22,2 %	55,60	28,68
Observed points	-	-	-	51,02	17,45

The first observation from the figures is the exceptionally high R^2 value of the aggregate model. It is caused by using the goals scored and conceded in the same season to model the results and scores of individual games. Since the goal difference of a team is one of the best determinants, if not the best, on the eventual amount of points cumulated, it is only natural that the covariance between these two variables is high. This in turn results in a high R^2 value.

Otherwise, the models derived from the expected goals method fare the best in this comparison on all fronts. Excluding the aggregate model's R^2 value, the models rank similarly on all model evaluation methods. The expected points achieves the best values while xG Poisson and expected goals models follow almost neck to neck behind. The shots on target and aggregate models are the fourth and fifth, respectively, and the predictive models fare the worse, as could be expected, since they cannot get data for their estimates from the matches.

5 Findings and conclusions

In this chapter the results of the thesis are further discussed and analysed. The chapter consists of three parts: First the findings of the research are discussed and explained. In part 5.1 the implications to research and practice are considered and critically evaluated. In the last part the limitations of this research are scrutinized and suggestions for future research in the field are made.

The research questions stated in chapter 1 were:

1) How accurately the expected goals model is able to explain match outcomes and success over a season?
2) Can the model's accuracy be improved with a Poisson distribution?

The ability of the expected goals model to explain match outcomes and success over a season was evaluated by benchmarking it to other models. First the match outcomes were analysed in chapter 4.1. The expected goals model was able to estimate the correct result on 57% and correct score on 17% of all instances. When compared to the benchmarks, only the xG Poisson model could do slightly better. While the numbers may seem as modest, it is important to remember that football is one of the most complex games and the complexity makes it immensely difficult to model. The low-scoring nature of the game causes a lot of variation and unpredictability to the results and score lines.

The model struggles to estimate correctly a few score lines. The figure 4.1 displays the over- and underestimations very informatively. Games resulting in one or both of the teams scoring zero goals were underrepresented in the model's estimates. While most of the times teams are able to generate enough goal expectancy to score at least one goal, they often fail to capitalize on these probabilities. On the other hand, the model overestimates the amount of 1-1, 2-1 and 1-2 scores.

One phenomenon that has been known in the field is the so-called home advantage (Boyko, Boyko, and Boyko 2007; Pollard 2008). There are few factors that cause this advantage. While the away team has to spend time and effort in travelling before the match, the home team can stick to their ordinary routines and prepare optimally for the incoming match. Other factors affecting this home advantage are, for example, the referee bias, home crowd support and psychological factors such as territoriality.

This effect can also be seen in the figures presented in this thesis, in table 4.3: The game ends on average 45% of the time in home team's victory, 25% in a draw and 30% in the

away team's victory, which are inline with the home advantage phenomenon. The expected goals model, however, estimates that up to 32% of the matches end in a draw, while the home team is expected to win in 44% of time. This means that out of all even matches, surprisingly many are converted into an away victory. In essence, the away teams can be seen to convert these tight matches in their favour and overperform their goal expectancy more often than the home team. This phenomenon was further demonstrated with the calculations of average error terms of both home and away teams in tight competitions.

When comparing the explanatory power of the different models on individual games, the expected goals model performed the best. The accuracy could not be improved with a Poisson distribution. When modelling success, i.e. the amount of points the teams are able to accumulate throughout one season, the expected points method fared the best. While also deemed as a biased model, the expected points could estimate the distribution of points most accurately. The xG Poisson model could improve the accuracy of the eventual estimated point distributions of the teams, even though the differences with the non-Poisson fitted model were minimal.

In practice this study highlights the fact that in order to estimate the power balances of each teams and the true performances of both sides, one cannot blindly trust the numbers. While the expected goals model does give valuable insight into the probabilities of the game and it depicts fairly accurately the events occurring in the field of play, it struggles to be reliable in estimating results and score lines systematically. The complexity and low-scoring nature of the game cause difficulties when trying to model individual matches, and the expected goals model is not an exception in this sense. In single matches, the sample sizes are too small for a model to effectively estimate the match outcome with precision. Additionally, when considering both team and player performance analyses, the model omits a lot of factors and implicit causal relationships of the game.

5.1 Limitations and future research

In this thesis only four leagues were used as a source of data: The English Premier League, Spanish La Liga, German Bundesliga and Italian Serie A. The leagues are fairly similar in terms of the level of teams and players as they are recognized as the best four leagues in the world. Thus, these leagues represent only the best football teams and players. Therefore, the qualitative spectrum of the teams in this thesis was quite narrow

and including other leagues could yield different results as the players would not be as effective in both converting the goal scoring opportunities into actual goals and, respectively, blocking opposition's efforts on goal.

In future research a broader set of leagues could be applied to the study to assess the explanatory power of the expected goals model. Another point to consider related to this is if there is a need to create some type of coefficient that could be applied to certain teams or players to get better results on the actual goalscoring probabilities. Logically, players who are more valuable and are playing in better teams should be better at converting goalscoring opportunities into goals than an average player. The current model, however, treats all the players the same, while in reality players differ in their capabilities.

The third and final suggestion for future research is the home advantage in even games. Researching the goal expectancies and converted goals in situations where both teams are vigorously trying to turn the game into their favour could yield some interesting results. There might be some underlying patterns to be discovered or, for example, behaviours that result in home team's poor performance in these even competitions.

References

Armstrong, J. Scott, and Fred Collopy. 1992. “Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons.” *International Journal of Forecasting* 8(1): 69–80.

Boyko, Ryan H., Adam R. Boyko, and Mark G. Boyko. 2007. “Referee Bias Contributes to Home Advantage in English Premiership Football.” *Journal of Sports Sciences* 25(11): 1185–94.

Brooks, Joel, Matthew Kerr, and John Guttag. 2016. “Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, , 49–55.

Cain, Michael, Law David, and David Peel. 2000. “The Favourite-Longshot Bias and Market Efficiency in UK Football Betting.” *Scottish Journal of Political Economy* 47(1): 25–36.

Chen, Chao, Jamie Twycross, and Jonathan M. Garibaldi. 2017. “A New Accuracy Measure Based on Bounded Relative Error for Time Series Forecasting.” *PLoS ONE* 12(3): 1–23.

Constantinou, Anthony Costa, Norman Elliott Fenton, and Martin Neil. 2013. “Profiting from an Inefficient Association Football Gambling Market: Prediction, Risk and Uncertainty Using Bayesian Networks.” *Knowledge-Based Systems* 50: 60–86. http://dx.doi.org/10.1016/j.knosys.2013.05.008.

Eggels, H P H. 2016. “Expected Goals in Soccer: Explaining Match Results Using Predictive Analytics.” Eindhoven University of Technology. https://pdfs.semanticscholar.org/f4aa/54ae207e3382b5bae1eae7131c2baf566d39.pdf?_ga=2.78249492.1017351469.1548261183-769100621.1548261183.

Elo, A. (1987). *The Rating of Chess Players, Past and Present*. *New York: Arco*. Retrieved from http://www.amazon.com/Rating-Chess-Players-Past-Present/dp/0668047216

Elofootball. (2020) Elo probabilities for seasons 2014-15 to 2018-19. Retrieved April 20, 2020, from http://elofootball.com/

Fernández, Javier, Luke Bornn, and Dan Cervone. 2019. “Decomposing the Immeasurable Sport: A Deep Learning Expected Possession Value Framework for Soccer.” *MIT Sloan Sports Analytics Conference*: 1–18.

FIFA. 2018. *Revision of the FIFA / Coca-Cola World Ranking*. Retrieved April 20, 2020, from https://resources.fifa.com/image/upload/revision-of-the-fifa-coca-cola-world-ranking.pdf?cloudid=fzltr4s8tz3v3vy0aq01

Flepp, Raphael, and Egon Franck. 2019. UZH Business Working Paper Series (ISSN 2296-0422) *The Role of Boards' Misperceptions in the Relation between Managerial Turnover and Performance: Evidence from European Football*. Zurich.

Football-data.co.uk. (2020) Market odds for matches for seasons 2014-15 to 2018-19. Retrieved April 16, 2020, from https://www.oddsportal.com/

Green, Sam. 2012. "Assessing the Performance of Premier League goalscorers." OptaPro Blog. Retrieved April 20, 2020, from https://www.optasportspro.com/news-analysis/assessing-the-performance-of-premier-league-goalscorers/

Heuer, A., C. Müller, and O. Rubner. 2010. “Soccer: Is Scoring Goals a Predictable Poissonian Process?” *EuroPhysics Letters* 89(3): 2–6.

Heuer, Andreas. 2020. “Identification of Relevant Performance Indicators in Round-Robin Tournaments.” : 1–15. http://arxiv.org/abs/2003.03774.

Hvattum, Lars Magnus, and Halvard Arntzen. 2010. “Using ELO Ratings for Match Result Prediction in Association Football.” *International Journal of Forecasting* 26(3): 460–70. http://dx.doi.org/10.1016/j.ijforecast.2009.10.002.

Jones, P. D., N. James, and S. D. Mellalieu. 2004. “Possession as a Performance Indicator in Soccer.” *International Journal of Performance Analysis in Sport* 4(1): 98–102.

Karlis, Dimitris, and Ioannis Ntzoufras. 2003. “Analysis of Sports Data by Using Bivariate Poisson Models.” *Journal of the Royal Statistical Society Series D (The Statistician)* 52(3): 381–93.

Krueger, James S., and Michael S. Lewis-Beck. 2006. “Goodness-of-Fit: R-Squared, SEE and ‘Best Practice.’” *The Political Methodologist* 15(1): 2–4.

Legates, David R., and Gregory J. McCabe. 1999. “Evaluating the Use of ‘goodness-of-

Fit' Measures in Hydrologic and Hydroclimatic Model Validation." *Water Resources Research* 35(1): 233–41.

Link, Daniel, Steffen Lang, and Philipp Seidenschwarz. 2016. "Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data." *PLoS ONE* 11(12): 1–17. http://dx.doi.org/10.1371/journal.pone.0168768.

Lucey, Patrick et al. 2014. "'Quality vs Quantity': Improved Shot Prediction in Soccer Using Strategic Features from Spatiotemporal Data." In *Proc. 8th Annual MIT Sloan Sports Analytics Conference*, , 1–9. http://www.sloansportsconference.com/?p=15790.

Macdonald, Brian. 2012. "An Expected Goals Model for Evaluating NHL Teams and Players." In *MIT Sloan Sports Analytics Conference 2012*, , 1–8.

Nielsen, Eirik S, and Sander F Sandøy. 2019. "Profiting from Football Betting Using Artificial Neural Networks." Norwegian University of Science and Technology.

Pollard, Richard. 2008. "Home Advantage in Football: A Current Review of an Unsolved Puzzle." *The Open Sports Sciences Journal* 1(1): 12–14.

Power, Paul, Hector Ruiz, Xinyu Wei, and Patrick Lucey. 2017. "'Not All Passes Are Created Equal:' Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Data." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, , 1605–13.

Rathke, Alex. 2017. "An Examination of Expected Goals and Shot Efficiency in Soccer." *Journal of Human Sport and Exercise* 12(Proc2).

Robberechts, Pieter. 2019. "Valuing the Art of Pressing." In *StatsBomb Innovation in Football Conference 2019*, Leuven, 11.

Severini, Thomas, A. 2012. Analytic Methods in Geomechanics *Analytic Methods in Sports*. CRC Press.

Spearman, William. 2018. "Beyond Expected Goals." In *MIT Sloan Sports Analytics Conference*, , 1–17.

Štrumbelj, Erik, and Petar Vračar. 2012. "Simulating a Basketball Match with a Homogeneous Markov Model and Forecasting the Outcome." *International Journal of Forecasting* 28(2): 532–42.

http://dx.doi.org/10.1016/j.ijforecast.2011.01.004.

Tenkanen, Santeri. 2019. “Rating National Hockey League Teams : The Predictive Power of Elo Rating Models in Ice Hockey.” Aalto University, School of Business.

Understat. (2020). Expected goals values for seasons 2014-15 to 2018-19. Retrieved March 21, 2020, from https://understat.com/

Willmott, Cort J., and Kenji Matsuura. 2005. “Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance.” *Climate Research* 30(1): 79–82.