CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

xG Statistics in Football Matches: Predictions and Betting

Bachelor's thesis

Author: Sebastian Černý
Study program: Economics and Finance
Supervisor: Mgr. Luboš Hanus, Ph.D.
Year of defense: 2024

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

During the preparation of this thesis, the author used ChatGPT to expand vocabulary and writing style and for assistance with LATEX. After using this tool, the author reviewed and edited the content as necessary and takes full responsibility for the content of the publication.

Prague, April 29, 2024

Sebastian Černý

Abstract

The thesis studies the effectiveness of betting on the results of football matches using Expected Goals (xG) statistics from two sources Understat and FootyStats. It evaluates the performance of 6 different models in seasons 2021/2022 and 2022/2023 across the five highest-ranked European leagues using binary logistic regression to predict two possible results, either the home team winning or away team not losing. For betting, several strategies are used based on existing literature. The results are compared to the model containing traditional variables used commonly for predictions in football based on relevant literature. Using both a combination of xG and traditional variables, and only xG variables the results suggest that xG variables are effective for predicting the outcome of football games. The model containing only Understat xG variables yielded 4.18% return on investment (ROI) when betting on every match in both seasons, which was 3.6% more than the model with traditional variables. For betting only on particular matches based on certain criteria, the combined models with both types of variables had the best results, reaching 10.87% ROI that again outperformed the model with traditional variables by approximately 4.5%.

JEL Classification	C10, C53, L83, Z29
Keywords	football, expected goals, betting, odds, prediction, logistic regression
Title	xG Statistics in Football Matches: Predictions and Betting

Abstrakt

Tato bakalářská práce se zabývá efektivitou sázení na výsledky fotbalových zápasů pomocí statistiky Očekávaných gólů (xG) ze dvou zdrojů: Understat a FootyStats. Hodnotí výsledky šesti různých modelů v sezónách 2021/2022 a 2022/2023 napříč pěti nejlépe umístěnými Evropskými ligami pomocí binární logistické regrese pro predikci dvou možných výsledků: výhra domácího týmu nebo neprohra venkovního týmu. Pro sázení bylo použito několik strategií na základě existující literatury. Výsledky jsou porovnány s modelem obsahujícím tradiční proměnné, které jsou běžně používané pro predikce ve fotbale na základě relevantní literatury. Za použití jak kombinace xG proměnných a tradičních proměnných, tak pouze xG proměnných, výsledky ukazují, že xG proměnné jsou efektivní pro predikování výsledků fotbalových zápasů. Model obsahující pouze Understat xG proměnné dosáhl 4.18% návratnosti investice (ROI) při sázení na každý zápas v obou sezónách, což bylo o 3.6% více než model s tradičními proměnnými. Při sázení pouze na vybrané zápasy za určitých podmínek, kombinované modely s oběma typy proměnných, měly nejlepší výsledky dosahující 10.87% ROI, což opět překonalo model s tradičními proměnnými o přibližně 4.5%.

Klasifikace JEL	C10, C53, L83, Z29
Klíčová slova	fotbal, očekávané góly, sázení, kurzy, predikce, logistická regrese
Název práce	xG statistika ve fotbalových zápasech: Předpovídání a sázky

Acknowledgments

I would like to thank Mgr. Luboš Hanus, Ph.D., my thesis supervisor, who has supplied me with valuable feedback every time I was seeking a helping hand. His remarks and recommendations meant a lot to me and I was always trying to take from him a piece of advice for future work. It was my great pleasure to cooperate with him. Moreover, I would like to thank my colleagues from IES, who motivated me to grow during the whole study program.

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record

Černý, Sebastian: *xG Statistics in Football Matches: Predictions and Betting.* Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2024, pages 99. Advisor: Mgr. Luboš Hanus, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

xG Expected Goals

ROI Return on Investment

xGA Expected Goals Against

xGD Expected Goals Difference

xA Expected Assists

xPts Expected Points

EGBA European Gaming and Betting Association

NFL National Football League

NBA National Basketball League

FIFA Fédération Internationale de Football Association

UEFA Union of European Football Associations

ECI Euro Club Index

ELO Elo Rating System

H2H Head-To-Head

AFC Athletic Football Club

FC Football Club

BC Bergamasca Calcio

MLE Maximum Likelihood Estimation

VIF Variance Inflation Factor

FIX Fixed Betting Strategy

PROP Proportional Betting Strategy

FER Fixed Expected Return Betting Strategy

KEL Kelly Criterion Betting Strategy

FIB Fibonacci Sequence Betting Strategy

ANN Artificial Neural Network

Chapter 1

Introduction

Since the introduction of statistics in sports, the possibility of data-driven predictions for the outcomes of future matches has suddenly appeared. In the beginning, it was only enthusiastic researchers, who wanted to predict the outcomes, but over time more people such as journalists and also bettors started being interested. As more data was collected and new statistical metrics were introduced, there were more options for deep mathematical and statistical analyses that could be applied to various sports. The number of bettings on sports has constantly increased over the last few years. Labrador & Vallejo-Achón (2020) conducted a study to identify and characterize the situation in which young students in Madrid participate in sports betting. From the sample of 735 young students from Madrid (aged 18-25), 42.6% admitted having placed a sports bet at least once in their life. Moreover, the study shows that most of the bets were made online and the main target was football, particularly betting on the winning team (favorite by bookmaker's odds). Similarly, Ricijas *et al.* (2016) mention in their study on predictors of adverse gambling that in their sample of 1330 male Croatian students from all three types of secondary education in Croatia (aged 14-21), the majority of them acknowledged being involved in sports betting. This increase is also supported by the statistics that are published by the European Gaming and Betting Association. Europe's gambling market revenue was around 100 billion euros in 2019 and reached 109 billion euros at the end of 2022 according to H2 Gambling Capital (2022). However, the increase would probably be higher if not for the pandemic situation caused by COVID-19, which resulted in many sporting events being canceled. Thus, the market revenue in 2020 and 2021 dropped to 82 and 88 billion euros. The Association revealed expected revenues in the upcoming years, which show a clear positive trend with the expected revenue by the end of 2027, reaching 134 billion euros. In 2022, the casino was Europe's most popular online gambling product by revenue, representing a 39% share of gross gaming revenue. However, sports and other types of betting were close behind, generating 13.6 billion euros and accounting for 35%. Comparatively, EGBA members generate the biggest share (46%) of their online revenues from sports betting.

Football in particular is considered to be the most popular sport on the planet. According to Veroutsos (2023), the global viewership of football is around 3.5 billion people ahead of cricket with a billion fewer fans around the world. In the article published by FIFA, the last Football World Cup that held place in Qatar by the end of 2022, the association claims that around 5 billion people engaged with the tournament content across an array of platforms and devices across the media universe. For the final game between Argentina and France, the match achieved a global reach of close to 1.5 billion viewers. As a consequence of such popularity, football has also the highest demand across the betting world. It is estimated that 1 billion pounds is bet on football every year in the UK and it represents approximately 46% of the whole betting market according to Gambling Commission UK (2022).

The betting market is created by bookmakers, who create the odds that bettors can later bet on. However, the main goal of bookmakers is to create such odds that the betting company is profitable. Thus, bookmakers usually conduct deep analysis for all their matches and create odds based on statistical analysis. Moreover, the profit of the company is ensured by the overround, also known as the book percentage. The overround is the amount by which the bookmaker's odds of a match exceeds the probability of one when converted to probabilities. Newall (2015) explains this in his study on how bookmakers make money, but generally, the higher the overround, the higher the expected profit for the bookmaker will be. This means that to be profitable from betting on sports, particularly football, the bettor has to overcome several obstacles. The most efficient way to beat a bookmaker in a longer period is to analyze the matches better than the ones who created the odds. One of the first researchers who created a model for predicting football matches was Maher (1982), who investigated the Poisson model based on past papers, using several parameters.

Still, in football, there are debates on which statistics measure the performance of the team in past matches the best. Many researchers such as Prasetio & Harlili (2016) use goals as a main measure of performance and create variables based on the goals scored and conceded. However, many matches have a different result than the one that would correspond to the actual game. For example, one team could have created a significant number of chances, but had a bad day and still lost to a team that created one chance but converted it into a goal. Actual goals scored do not take this fact into account. That is why different statistics started to be calculated such as ball possession, number of shots, number of corners, etc. Nevertheless, these statistics had the same problem as goals, because there were matches, where the team had high ball possession, and a high number of shots and still lost. This is why the metric called Expected Goals (xG) was created. As the origin of this metric is not entirely certain, the whole concept started to gain some attention in the early 2010s. The early developments and adoptions of xG were driven primarily by the analytics community with blogs, forums, and independent researchers. In the article from Whitmore (2023), he mentions that the xG metric as we know it today, was introduced in 2012 by an analyst called Sam Green, who works for the analytical company Opta Analyst[1]. This metric measures the quality of a chance by calculating the likelihood that it will be scored by using information on similar shots in the past. As it is one of the first advanced metrics to become widely known among general football fans, the battle between the traditional way of viewing the game and the analytical way has begun. Many experts and ex-players criticize the xG metric, arguing it fails to accurately reflect performance since football ultimately depends on scoring goals. Even though this statement is true, studies (Maurath (2020)) show that the xG models can reach up to 92% accuracy when comparing the actual goals scored and xG values within a 95% confidence interval.

As justified above, betting has been on the rise along with the xG statistics newly introduced to the football public. This study aims to perform a detailed analysis of the xG statistics and its influence on predicting football matches in both a statistical and mainly a betting manner. The main goal is to determine whether the xG statistics is a great component for betting and if it is possible to be profitable while betting using the xG statistics. The thesis is structured as follows. In Chapter 2, we compare how the existing literature dealt with predicting sports in general and then in football. Which mathematical and statistical models were used and with what accuracy and results. Moreover, the chapter contains a literature review of betting strategies regarding the management of wealth and the methods that are used for both stock investment and sports betting. Chapter 3 will cover data collection, variable creation for pre-

[1] https://theanalyst.com/eu/

dictive models, a detailed analysis of the xG statistics as the main variable, and descriptive statistics of all variables with relevant tables. Chapter 4 discusses the use of logistic regression and its assumptions along with models that have been modeled for predictions. Chapter 5 is devoted to betting, it introduces the reader to how betting works and explains the betting strategies used in this work. Lastly, Chapter 6 is dedicated to results, showing the accuracies of particular models, their efficiency when combined with betting strategies from Chapter 5 using the return on investment (ROI) as the main criteria and graphical visualization of the profit/loss for each model.

Chapter 2

Literature Review

2.1 Predictions of Sports Matches in General

The purpose of this literature review is mainly to introduce the origins of predicting sporting events in general and primarily football matches. Moreover, to explain how and which betting strategies were created and similar to predictions, mention studies that have already tried betting on football. The literature review mentions several studies with a variety of methods used for both predictions and betting. Additionally, in Section 2.2 is mentioned the only academic study on the xG statistics and its predictive and betting power.

According to a study by Singla & Singh (2020), machine learning is used extensively in sports prediction. They also tested a few methods themselves when predicting the outcomes of the most popular football leagues in the world. In their study, three different methods are tested and compared. The Naive Bayes Classifier, Support Vector Machines, and Logistic Regression. After normalization, the logistic regression outperformed the other two methods with a maximum accuracy of 61%. The overall accuracy in their work was 53%. The claim that machine learning is the most popular technique when predicting outcomes in sports is supported by another study from Koseler & Stephan (2018), which performed a systematic literature review of machine learning applications in baseball analytics. Purucker (1996) used the Neural network method to predict 5 weeks of the National Football League (NFL) in season 1994 with an average accuracy of 60.7%. Kahn (2003) tried a similar approach also using the Neural network method for two game weeks in the 2003 NFL season with an accuracy of 75%. Another study this time for predicting the first 650 games of the basketball season in 2007 was conducted by Loeffelholz *et al.* (2009) using again the Neural network method. He trains a variety of neural networks such as feed-forward, radial basis, probabilistic, and generalized regression neural networks. He then compared his results to predictions made by numerous experts in the field of basketball. The best networks were able to correctly predict the winning team 74.33% of the time. There are also methods like Random forest, and XGBoost that are used in the study by Elfrink (2018) along with Logistic regression and GLM for predicting the outcome of baseball matches using historical data. He achieved the highest accuracy of 55.52% when using the XGBoost algorithm, but mentioned that better results can be achieved by using more data and better feature engineering. A different approach to predicting results was used by Hsu (2020) for the American Football Games by adding the candlestick charts compiled based on betting market data and using their behavior as one of the features in the predictive model. He then uses methods such as Neural Networks, Ensemble Learning, or the Random Subspace method for regression. The latter achieves the best results with an accuracy of 68.4%. More advanced methods were introduced in the paper from Caliwag *et al.* (2018), where it presents an improved technique for predicting basketball game results by implementing the cascading algorithm. The researchers combined Naive Bayes, Four Factor Analysis, and Fuzzy Logic Algorithms to predict basketball games in NBA season 2015/16 with 69% - 70% accuracy.

2.2 Predictions in Football

As football is the most popular sport globally, it is also one of the most studied sports. Football is a very complicated thing to analyze as there are hundreds of factors that affect the outcome of each match. Firstly, it is the basic ones such as the strength of the team, position in the league, number of goals scored in the league so far, number of conceded goals, number of injured players, etc. However, if we want an exact prediction, we would ideally want to include every factor, even the less common ones such as weather conditions, size of the stadium, rest days between the matches, etc. As in the general predictions in sports, many methods can be used to predict the outcomes of football matches.

One of the first analyses to predict football matches was conducted by Maher (1982), who investigated the Poisson model that was rejected by previous authors in favor of the Negative Binomial which was tested also by Pollard (1985) for predicting the number of goals scored. Maher used parameters representing the team's inherent attacking and defensive strength and found the most appropriate model from a hierarchy of models. Concluding that the independent Poisson model showed that the team qualities had significant effects on match outcomes. The model provided a reasonably good fit to the observed score distributions in most cases. However, he stated that the model did not fit well for the differences between scores, which led him to introduce a bivariate model, which accounts for the correlation of 0.2 between scores of home and away teams, which improved the fit significantly for score differences. A few years later Dixon & Coles (1997) made a few changes to Maher's model. Their model accounts for the lack of independence between events in a single football match. For example, if one team scores, it affects the strategy of both teams for the rest of the match. Also, they adjusted the model so that low-scoring draws are more probable and results with only one goal are less likely to occur. Lastly, Dixon and Coles use bookmaker's odds as independent variables and conclude that their model is shown to have a positive return when used as the basis of a betting strategy.

As predicting football outcomes is quite challenging, machine learning methods are extensively used in this context. Hucaljuk & Rakipovic (2011) compare six different methods in their paper including Naive Bayes, Bayesian Networks, LogitBoost, The k-nearest neighbors algorithm, Random Forest, and Artificial Neural Network. The best results were achieved by Artificial Neural Network (ANN) with a prediction accuracy of up to 68%. The authors stated that there is room for further improvement such as including the form for every player and larger data set for extensive learning.

There were also some studies where researchers predicted the results of a single team across one or several seasons. Joseph *et al.* (2006) looked in their paper at the performance of an expert-constructed Bayesian Network compared with other machine learning methods for predicting the outcome of matches played by Tottenham Hotspur in 1995-1997. The machine learning techniques were an MC4 decision tree learner, a Naive Bayesian learner, a Data-Driven Bayesian, and a k-nearest neighbor learner. The accuracy of the expert-constructed Bayesian Network outperformed the others with an accuracy of 59.21%. A similar thing was tested by Owramipur *et al.* (2013), but this time for a Spanish team FC Barcelona in the season 2008/09. However, they added several additional variables to their model such as the psychological state of the team, weather conditions, and the absence of key players. The accuracy of this model reaches 92%, but the limitations of predicting only the results of one team and in this case, the team that won the league this season are high as the data are imbalanced.

One of the most effective and often-used methods is the logistic regression. Buursma (2010) tested the logistic regression to predict 15 years of Dutch football. Even though he reaches only 55% accuracy, in his research he manages to achieve profitable results when using the right betting strategy. Another study using logistic regression was conducted by Prasetio & Harlili (2016) for predicting matches of the Premier League season in 2015/16. They train their model on 5 previous seasons and reach an accuracy of 69.5%. Interestingly, their research uses only significant variables from several different research papers, which are the defensive and attacking strengths of both home and away teams. They reach an accuracy of 69% using only the defensive strength of both teams. Moreover, Chang *et al.* (2023) use logistic regression as well in their study among others, which are Random Forest, and Gradient Boosting Decision Tree, which is an additional model based on the idea of boosting integrated learning. They selected the English Premier League and Spanish La Liga in the 2018-2022 seasons. The Random Forest model was the most accurate with an accuracy of 66.7% on the training set and 63.8% on the test set. The logistic regression model had only 59% accuracy. Lastly, one of the main questions of van Wijk (2021) master's thesis was which machine learning is the best for predicting the outcome of football matches. He gathered data from nineteen seasons of the English Premier League from 2002/03 to 2020/21 and tested three different methods, namely Logistic Regression, Neural Network, and Random forest. By the end of his research, he concluded that the model with the highest accuracy is the Logistic Regression.

Finally, Partida *et al.* (2021) focuses directly on the xG statistics and its application for predicting football matches. Their study aimed to examine the predictive capabilities of mathematical models based only on the expected goals statistics. They created 3 different models that use xG statistics as input for the Poisson distribution to generate goal-scoring probabilities that were combined into probabilities of home win, away win, or draw outcomes. Afterward, the sample probability matrix for each model is created to calculate the probability of each possible outcome. The overall accuracies for all 3 models were around 50%. However, when betting only on match outcomes with a probability greater than 70%, two models managed to be profitable. The study has several limitations such as not including other contributing factors such as the team's defensive abilities in a more meaningful way or the current state of the league. It is suggested that implementing these changes would require more advanced modeling approaches such as regression and that was beyond the goals of their study.

2.3 Betting in Football and in General

When betting on sports, there are two main components, the first one is creating the predictive models that give the bettor probabilities for the given match outcomes, and the second one is the betting strategy and money management techniques to optimize the profitability of a particular model and its predictions. Betting on sports also requires a comprehensive understanding of the market such as the meaning of odds, and the bookmaker's margin, but also risk management as in betting, the loss of wealth can be remarkably quick.

As Uhrín *et al.* (2021) mention in their study, the first two most notable approaches to allocating wealth were introduced by Markowitz (1952) with his concept of balancing the return and risk of a portfolio and for betting more notable Kelly Criterion introduced by Kelly Jr. (1956) to maximize the long-term growth in a scenario where the same game is being played repeatedly.

Thus, the use of the Kelly strategy when betting was thoroughly examined by many researchers in the following years. A study from Smoczynski & Tomkins (2010), creates an explicit formula for the optimal strategy for betting allocation on horse races, which resolves the strategy that Kelly proposed, which is believed to be incomplete by the authors. Additionally, Grant & Buchen (2012) studied the problem of betting on multiple simultaneous games, concluding that using the Kelly strategy on multi-bets of all levels outperforms the portfolio optimization approach of betting on single outcomes only.

However, in sports, the majority of bettors prefer to look for value bets (outcomes with positive expected value). Consequently, finding the right investment technique is somewhat overlooked in favor of simpler methods, such as ones exploiting the sports-betting market using machine learning Hubáček *et al.* (2019). For football in particular, van Wijk (2021) provides significant insight into the betting strategies for football. His study was conducted mainly to answer the question of which machine learning method is the best for predicting the outcomes of football matches. Moreover, one of the subquestions was to find out which betting strategy can create opportunities for profitable results. He studied six different money management techniques, including the Kelly Criterion for betting. Among these advanced techniques are also the Fibonacci sequence technique and the variance-adjusted technique. The latter is a simplification of Markowitz's portfolio management and it was introduced by Rue & Salvesen (2001). Van Wijk also tried popular strategies among bettors, which are fixed betting, proportional betting, and lastly fixed expected return betting. In combination with the logistic regression, he concluded that the Fibonacci sequence was performing the best with an average of roughly 15% ROI. However, he stated that this method is riskier than other techniques and in reality, can lead to large stakes, which is not very rational. Therefore, he compared the results of fixed betting and proportional betting as the following two best-performing techniques. The proportional technique showed results of 4.84% ROI and the fixed 5.86% ROI. So the answer to his subquestion was that Fixed Betting is the best-performing betting strategy when compared to several others.

This study aims to continue in what Partida *et al.* (2021) started and use the logistic regression as a more advanced modeling technique proposed by the authors themselves to contribute to the existing literature. The use of logistic regression is justified in the literature review as many authors achieved positive results using this technique. Particularly, binary logistic regression is used to estimate the probabilities of two possible outcomes in a football match (the home team winning against the away team not losing the match), which are then compared to implied probabilities from bookmaker's odds and used for betting purposes. Several models are created from variables based on the existing literature and new xG variables explained in Section 3.3.3. The use of betting strategies is inspired by van Wijk (2021) and Hubáček *et al.* (2019).

Chapter 3

Data & xG Statistics

3.1 Data Collection

The data used in this research cover 25 seasons, 5 years for 5 leagues. The leagues are the English Premier League, Italian Serie A, French Ligue 1, Spanish La Liga, and German Bundesliga, which are the five highest-ranked leagues based on the UEFA coefficient[1]. Each season in the Premier League, La Liga, Serie A, and Ligue 1 has 380 matches as 20 teams play against each other at home and away once. However, in the 2019/20 season, there were only 279 matches in Ligue 1 as the season was abandoned due to the COVID restrictions in the country that prohibited all sporting events. Bundesliga has only 18 teams in the league so the number of matches in one season is 306. The main source of data was Football-Data.co.uk[2], as they provide datasets of each season with all matches, including variables such as the date of the match, the name of the home team, the name of the away team, the number of goals scored by both teams, the match result along with the traditional match statistics such as number of shots, corners, yellow and red cards. The xG data were gathered from FootyStats[3] and Understat[4] as these are the only two sites that provide xG for all matches from the five highest-ranked leagues for several historical seasons. However, all sides provide only match statistics that happened in the particular match. For this reason, pre-match variables had to be calculated from the match statistics before the given match. The betting odds were also obtained from Football-Data.co.uk, as they include several betting sites and their odds

[1] www.uefa.com/nationalassociations/uefarankings
[2] https://www.football-data.co.uk/data.php
[3] https://footystats.org/
[4] https://understat.com/

for every match. Additionally, they provide also maximal closing odds, which are the maximal odds before the match on the market. Moreover, from Euro Club Index[5] (ECI), we gathered the ECI variable, which is a ranking of the football teams in the highest divisions of all European countries showing their relative playing strengths at a given point in time. The ECI values are derived from historical and actual sporting results in league matches, national cup matches, and UEFA Champions League, UEFA Europa League, UEFA Conference League, and UEFA Super Cup matches. I gathered the values for each team in the dataset at the beginning of each season.

3.2 Creation of Traditional Pre-Match Variables

Firstly, as datasets from Football-Data.co.uk are in a seasonal format, I had to merge them. Additionally, I added two variables from the datasets that I gathered from FootyStats as they included the number of the game week and the timestamp of each match. The timestamp variable was important later when dealing with postponed matches.

The result variable had three possible values, H representing the home team winning, A representing the away team winning, and lastly D if the match resulted in a draw. However, my study aimed to use binary logistic regression, for which the outcome variable has to be binary (Wooldridge (2012)). Thus, I defined the result variable as follows:

$$FTR = \begin{cases} 1, & \text{if Home team wins} \\ 0, & \text{if Away team not lose} \end{cases}$$

Furthermore, new pre-match variables had to be created so that they could be used for making predictions for the given match. As a result, all of the variables mentioned below are calculated up to each match. The goal of my thesis is to compare the models and find out if the xG statistics have a significant impact on betting in football, therefore the traditional variables are created based on the existing literature, mostly van Wijk (2021). Based on the match results, the average number of points was calculated for both teams by dividing the number of points the team had by the number of matches played so far. Additionally, the cumulative goals scored and conceded were calculated as the sum of goals from all previous matches in the season. Then goal difference is calculated,

[5] https://www.euroclubindex.com/

which is important for determining the league position. It is calculated by subtracting the team's number of goals conceded from the goals scored, which means that a positive goal difference is better. The criteria based on which the position is concluded for all five leagues are as follows: 1) the number of points a team has, 2) a higher goal difference, and 3) a higher number of goals scored. Then the cumulative goal difference is divided by the number of matches played so far so that the scale of all variables stays the same during the whole season. The attacking strength of both teams is calculated by taking the average number of goals scored by the team and dividing it by the average number of goals scored in the whole league. Similarly, the defensive weakness is calculated by taking the average number of goals conceded by a team and dividing it by the average number of goals conceded by the whole league. Higher values of the attacking strength indicate that the team is performing well offensively compared to the rest of the league. On the other hand, well-defending teams have lower values of defensive weakness. Furthermore, the form variable of both teams is created so that it captures their recent performance. It is calculated by taking the number of points obtained in the last 5 (or less if it is the start of the season) matches by the team and dividing it by the maximum number of points the team could have gotten. This approach ensures that for the first 5 matches in the season where the team has not played 5 matches yet, the scale of the variable is the same. For the first game week in each season, this variable is set to zero as it does not make sense to consider the form from the previous season because the team has not played for more than 2 months and more importantly the strength of a team might have changed due to possible transfers in or out of the club during the summer break. Also, this variable is multiplied by the ECI to take into account the strength of the opponents faced in the recent matches. The ECI was adjusted by taking the original value from the Euro Club Index and dividing it by the average value of the whole league to highlight the comparative strength of a team against the rest of the league. By adjusting the form with ECI, we make sure that the form is as accurate as possible since playing against 5 difficult opponents is more challenging than facing teams from the bottom of the league. Home team win ratio and away team not losing ratios are calculated to capture the effect of playing at home and away. These ratios are calculated by taking the number of wins of a home team when playing at home and the number of draws and wins of an away team when playing away and dividing it by the number of games played either at home or away. Furthermore, the H2H ratios are calculated as historically some games are more specific (for example town rivalries), and the results are often different than expected. Due to the importance of the match, the teams tend to approach the match more seriously than others and the performance in the rest of the season is less important and the chances of both teams to win the match are more even. The H2H ratio is calculated by looking at the results of the team's last five encounters and comparing the wins. Rather than incorporating draws into this calculation, it exclusively considers the win/loss record. Specifically, it computes the ratio by aggregating the number of wins achieved by both the home and away teams and then dividing by five. This method ensures the ratio remains within a range of 0 to 1 during the whole season. For example, if the home team won 4 of the last 5 matches and one game resulted in a draw, the value for the home team will be 0.8 and 0 for the away team. If the team got promoted to the first league and has not played any matches against the other team, the value will be 0 for both teams. Taking only the H2H ratio from the last 5 matches ensures at least partial relevance for the prediction as the results are still relatively recent and the strength of both teams should not be significantly different. The last traditional variables calculated are the average number of shots on target, and the average number of corners and fouls. This is done by calculating the cumulative number of shots on target, corners, and fouls in the season so far and dividing it by the number of matches played. For the first matches of each season, the variables are set to 0 as it does not make sense to involve the values from the previous season. The average summer break is around 2 months and between the seasons, the changes in the quality of teams are the most significant.

3.3 Expected Goals (xG)

3.3.1 Explanation of the Metric

Expected goals (xG) is a metric designed to measure the probability of a shot resulting in a goal. An xG model uses historical information from thousands of shots with similar characteristics to estimate the likelihood of a goal on a scale between 0 and 1. For example, a shot with an xG value of 0.2 is one that we would generally expect to be converted twice in every 10 attempts.[6] It is important to understand that each xG model has its characteristics and the values for different models might differ, which will be illustrated in Section 3.3.2.

[6] https://statsbomb.com/soccer-metrics/expected-goals-xg-explained/

However, the majority of xG models do share the main factors such as distance to the goal, angle to goal, body part with which the shot was taken, type of assist or previous action, and others. After the match is played, the sum of all xG values for each team is calculated and the better team in the match is defined as the one with a higher value of xG by the end of the match (as the team was expected to score more goals). According to a study from Maurath (2020), depending on the quality of the xG model, between 79% and 93% of team seasons should be expected to match xG to actual goals within a 95% confidence interval. This research will use two different sources for the xG values: FootyStats and Understat, each with its own xG model. In the following Section 3.3.2, both sources will be analyzed and the main differences will be highlighted.

3.3.2 In-Depth Analysis of the xG from Both Sources

As mentioned above, we have two sources of xG statistics as the values differ across models. In the following section, the interesting parts and the differences between the FootyStats model and the Understat model will be highlighted.

Statistic	FootyHomexG	FootyAwayxG	UnderHomexG	UnderAwayxG
Min.	0.110	0.010	0.000	0.000
1st Qu.	1.260	1.040	0.870	0.640
Median	1.580	1.340	1.390	1.110
Mean	1.635	1.395	1.548	1.262
3rd Qu.	1.960	1.690	2.070	1.710
Max.	4.450	3.910	6.880	6.500

Table 3.1: Summary Statistics for Expected Goals (xG)

Summary Table 3.1 represents the distribution of xG values in the dataset from both sources. FootyHomexG represents xG achieved by a home team in a match and FootyAwayxG represents the value obtained by the away team (the same for Understat). The first difference that the summary table highlights is the range of values, the Understat xG values have higher variability. The difference between the highest and the smallest value that a home team achieved is 6.88 while for FootyStats the difference is around 4.34. This would indicate that Understat might have some outliers and it has in fact, however in football, there are outliers in the results as well. Typically, at least once in a season, a match ends with an unusually high score. In every season of my dataset, there was a game in which one team scored at least 7 goals, which is higher than the maximal value of xG represented in the dataset. Moreover, the minimum value for both home and away matches is not 0 for the FootyStats model, while Understat has 0 present six times in the dataset (4x for an away team, 2x for a home team). Looking at the mean of both models, the mean is higher for the team playing home, which is logical as statistically home teams win more matches and also create more chances implying higher values of xG. For the Understat model, the difference is around 0.29, while for FootyStats it is slightly less around 0.24.

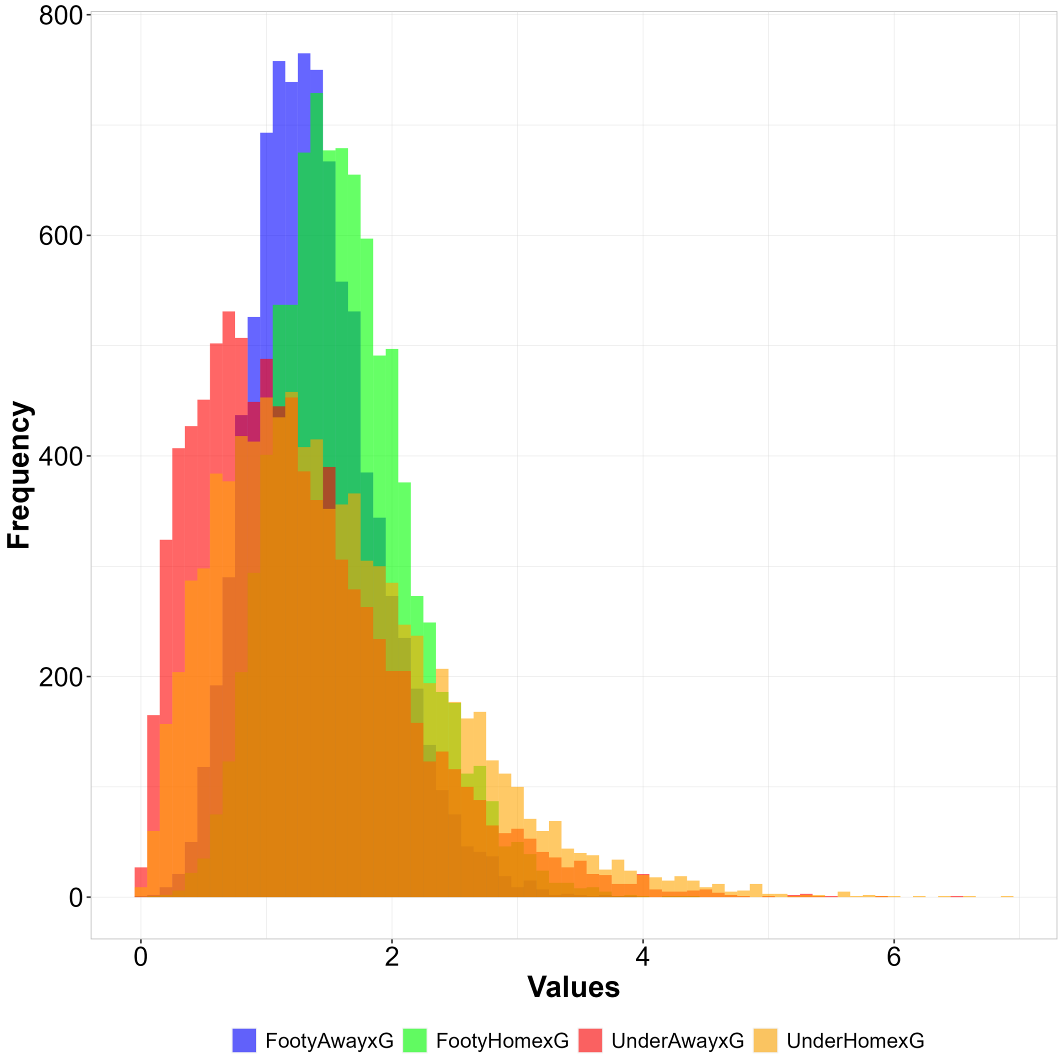

Figure 3.1: Distribution of xG Values

When looking at Figure 3.1 representing the distribution of each variable, we can see that the distribution is right-skewed for both models with a tail extending more towards higher xG values. This is logical as we cannot have negative values for xG and it has no upper limit. The graph confirms that the Understat model tends towards higher values of xG. The distribution peaks suggest that on average, the FootyStats model predicts a slightly higher number of expected goals than the Understat one, which is supported by the fact that FootyStats xG has a higher mean than Understat xG. On the other hand, there is a significant overlap of distributions, which means that many times both models have similar values of xG.

Correlation	
Home Games	**Away Games**
0.6312	0.6518

Table 3.2: Correlation between FootyStats and Understat xG

Table 3.2 shows a correlation between these two variables suggesting a positive relationship between them. The scatter plots in Figure 3.2 support this, with a cluster of data points forming an upward trend, implying that as one model's xG increases, so does the other. The trend is more concentrated along the line of best fit, indicating consistency in the relationship. However, they are not extremely close to 1, which means that there are still differences between the two models and how they calculate xG.

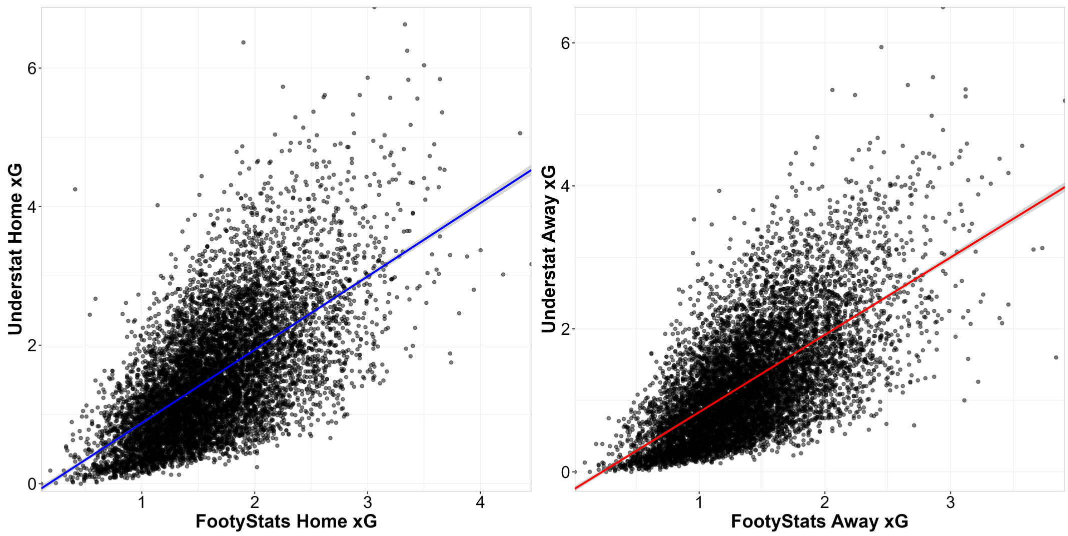

Figure 3.2: Correlation between Home and Away xG

Furthermore, we calculate the difference between the xG values and the actual goals scored to highlight how the predicted expected goals differ from reality. It is done by subtracting the xG variables from actual goals.

	FootyHome	UnderHome	FootyAway	UnderAway
Min.	0.0000	0.0000	0.0000	0.0000
1st Qu.	0.3800	0.3200	0.3600	0.2800
Median	0.8100	0.6500	0.7800	0.5600
Mean	0.9645	0.7935	0.8734	0.7084
3rd Qu.	1.3800	1.1000	1.2300	0.9800
Max.	6.1700	5.0600	5.6100	4.9300

Table 3.3: Absolute Difference between Actual Goals and xG

Firstly, Table 3.3 represents the absolute difference between the values. We can see that for both home and away matches the Understat values of xG have smaller means, which indicates that the model might be more precise than the one from FootyStats. For both models, the minimal values are 0, which means that there were matches where the team scored the exact number of goals that was expected by both models. This was the case for the home team in 34 matches for both models and for the away team in 31 matches for the FootyStats model and 37 for the Understat model. Moreover, the means are higher for home teams, which might again be due to the nature of football matches, where the home team usually creates more chances than the away team. This means that the model will predict higher values of expected goals due to the number of occasions a team could have scored. However, in football, the team does not convert all chances into a goal and thus the difference is created. As explained above, there are matches, where the team scores 7 or more goals, which on a professional level happens only when the team scores even from chances that were not very likely to be converted into a goal. If we find what match has this maximal difference it is the match from 2022, where Liverpool FC scored 9 goals against AFC Bournemouth.

Date	Home	Away	FTHG	FootyxG	UnderxG
27/08/2022	Liverpool	Bournemouth	9	2.83	4.86

Table 3.4: Liverpool - Bournemouth Match With xG Values

In Table 3.4, we see that the difference is very high and when compared to the xG represented by the Understat model, we are referring to a 2.03 difference between the two models. Nevertheless, the difference between the xG of the Understat model and the actual goals is still 4.14. Still, it supports the claim that the Understat model on average inclines more towards more extreme values than FootyStats.

Statistic	**FootyHome**	**UnderHome**	**FootyAway**	**UnderAway**
Min.	-6.1700	-4.8100	-5.6100	-4.9300
1st Qu.	-0.6000	-0.5900	-0.5000	-0.5200
Median	0.2400	0.1000	0.2700	0.1200
Mean	0.1015	0.01384	0.1453	0.01282
3rd Qu.	0.9600	0.6900	0.9400	0.5900
Max.	4.3500	5.0600	3.2200	4.5000

Table 3.5: Difference between Actual Goals and xG

As we examined the absolute differences between goals and xG, we now take a look at the difference, which is the actual number of goals scored subtracted from the xG value, but this time not in absolute values. This way, we can see also the direction in which the difference is. Particularly, negative values indicate that more goals were scored than expected, and the opposite for positive values. Negative values of the first quartile for all 4 variables in Table 3.5 suggest that in at least 25% of the matches, there were more goals scored than expected. Looking at the mean and median, we can conclude that both models tend to slightly overestimate the expected goals against the actual goals for both home and away teams as the sign is positive, but the means are nearly 0, suggesting that both models are, on average, highly precise. Additionally, the Understat model seems to be more accurate as the mean is closer to 0. However, the Understat model has the biggest difference with a positive sign (meaning that the team scored fewer goals than expected by the model) for both home and away teams. This corresponds to the distribution of Understat xG values that are more skewed to the right and also the overall tendency of the model to estimate high values of xG (the long tail to the right).

Date	**Home**	**Away**	**FTHG**	**UnderxG**	**FootyxG**
15/04/2019	Atalanta	Empoli	0	5.06	4.35

Table 3.6: Atalanta - Empoli Match with Home Team xG Values

In the particular match in Serie A between Atalanta BC and Empoci FC in 2019, represented by Table 3.6, the Understat model expected 5.06 goals for the home team even though they did not score a single goal. However, the FootyStats model also expected 4.35 goals scored from Atalanta in the match and they had 47 shots in the match, which was the most shots in the European five-highest-ranked leagues in a single match since the 2006/2007 season.[7] So this match was rather a combination of underperformance from the Atalanta offense in combination with bad luck. Moreover, if we examine the 3rd quartile, we can see that the Understat model has a smaller value than FootyStats for both the home and away teams. This indicates that the FootyStats model tends to expect more goals than the actual goals scored more often than Understat. At the same time, Understat might be more accurate for most of the games. Still, the variability of values might cause more significant errors for a smaller portion of games where considerably fewer goals were scored compared to the expected values.

Threshold	**Understat (%)**	**FootyStats (%)**
0.00	70.48	59.43
0.10	71.47	60.21
0.25	72.93	61.90
0.50	73.35	63.07
0.75	73.37	62.26
1.00	72.21	61.19

Table 3.7: Comparison of Correctly Estimated Matches

To create an example of how accurate the metric is, a table with 6 thresholds has been created, illustrating the ability to correctly address the performance of both teams in the match. In an ideal world, for this metric to perfectly address a team's performance in a match, the accuracy would be 100%. These thresholds had been set, because it is not possible to directly say that if a team had more expected goals in a match by 0.01, they were automatically a better team and should have won the match. The calculation for the percentage of correctly estimated matches is by subtracting the xG value that the away team obtained in a match from the xG value a home team had. If the result is higher than the threshold, the match is estimated to be a home team win, otherwise, the result is an away team not losing. This estimated result is then compared to the actual result of the match and the percentage of correctly estimated matches is calculated. Looking at Table 3.7, we see that again the Understat model has better results. The best threshold for FootyStats is 0.5 with approximately 63.1% accuracy. For the Understat model, the best threshold seems to be 0.5 and 0.75 with an accuracy of around 73.35%, which is logical

[7]https://www.bbc.com/sport/football/47943076

as Understat has higher variability in the values. To sum it up, with the ideal threshold, the expected result based on the Understat xG model is in 73.37% of all 9029 matches the same as the actual result. Even though for FootyStats, the agreement prevails in only 63.07% of matches, still it demonstrates the ability of this metric to correctly capture the performance shown on the field in the majority of matches.

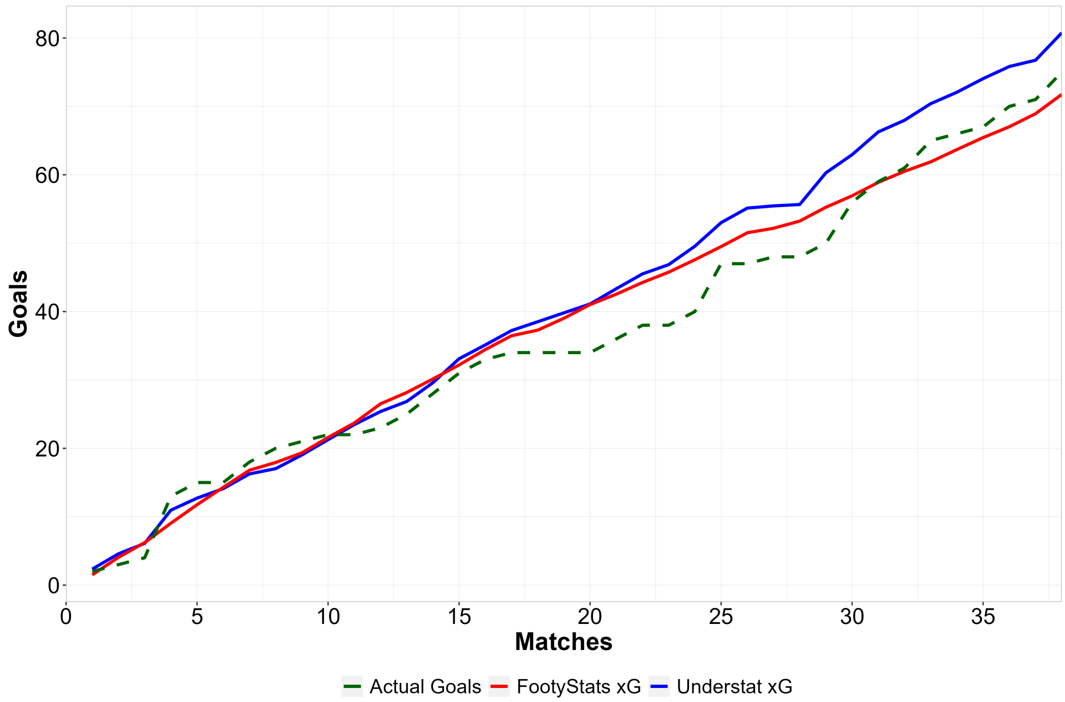

Figure 3.3: Cumulative xG and Actual Goals for Liverpool per Match (2022/2023 Season).

Lastly, we can compare the development of expected goals and goals scored during the season. In Figure 3.3, we examine the performance of Liverpool in the 2022/2023 season as they are considered to be one of the most offensive teams in Europe. Firstly, we look at the expectancy of goals by both models is approximately the same up until the second half of the season, where the Understat model started to expect slightly more goals than FootyStats. The most significant divergence between them happens at the end of the season. After the last match of the season, the peak difference between both models reaches exactly 9 expected goals. Secondly, we can see that in the first half of the season, Liverpool scored more or less the number of goals that they were expected to. Around the half of the season, they started to underperform according to both models and scored less than predicted. Compared to the FootyStats model, the biggest difference was 7.76 goals after the 23rd match in the season against Crystal Palace FC, which resulted in a 0-0 draw and Liverpool should have scored 1.51 goals based on the xG model. The biggest disagreement was even slightly higher in comparison with the Understat model, which reached a 10.3 difference between the expected and actual goals scored after the 29th match against Arsenal FC, which resulted in a 2-2 draw. According to the model, Liverpool was supposed to score 4.64 goals in the match. However, by the end of the season, the actual number of goals Liverpool scored (75) was higher than expected by the FootyStats model, which expected them to score 71.75 goals. Based on the Understat model, Liverpool still underperformed a little as Understat expected 80.75 goals scored by the end of the season.

To sum it up, we can conclude that there is negligible ability from both models to correctly estimate how many goals teams should score, which directly implies the ability to capture the performance of both teams that can later serve as a quality prediction component. From most of the results of our brief analysis of xG values, we see that the Understat model seems to be more precise than the FootyStats one. However, we will use both xG models for our study, as we are interested in their ability to predict football results.

3.3.3 xG Variables

For the xG statistics, pre-match variables had to be created as well. Firstly, the average xG and xGA (xG against the team) are calculated by dividing the cumulative xG in the season by the number of matches played so far. These averages are then used for calculating the similar variables as in Section 3.2 such as the attacking strength and defensive weakness of both teams, but this time the attacking strength is calculated by taking the average xG of the team divided by the average xG of the whole league. The defensive weakness is calculated by taking the average xGA of one team and dividing it by the average xGA of the whole league. Similar to the original attacking strength and defensive weakness, higher values of attacking strength show that the team is creating more opportunities than the rest of the league and lower values of defensive weakness indicate that the team is defending well as there are fewer chances created against them compared to the rest of the league. In addition, the attacking efficiency and defensive efficiency are calculated for both teams. The attacking efficiency is calculated by taking the total number of goals scored in the season and dividing it by the cumulative xG obtained in the season so far. The same calculations are done for the defensive efficiency but with the total number of goals conceded and divided by cumulative xGA. Moreover, the xG form variable is also calculated for both teams by taking the xG from the last 5 games (or less if the team has not played that many matches in the season) and dividing it by the 5 (or less). This variable is then multiplied the same way as the original form variable by the scaled ECI variable to account properly for the performance in the last 5 matches as creating great chances is more difficult against stronger opponents. Moreover, the xG per shot variable is created, which is calculated by taking the cumulative xG in the season so far and divided by the total number of shots. Lastly, the league position based on xGD (xG - xGA) is calculated for all teams with only one criterion, which is the value of xGD. Higher values of xGD imply a better league position.

3.4 Descriptive Statistics

The final dataset consists of 9029 matches across 5 seasons and 5 leagues including pre-match variables calculated in a way, that they can be used for making predictions. The dataset includes 66 variables including three non-numerical variables: Div, which is a variable indicating from which league the match is, and the names of the home team and the away team. The distribution of match results is depicted in Table 3.8. Overall, in all 5 seasons, 56.6% of matches resulted in an away team not losing and 43.4% in the home team winning the match. Apart from season 2 (2019/2020), where the French league was abandoned in April 2020 due to COVID-19 restrictions, there were 1826 matches in each season across all 5 leagues. The highest result discrepancy was in season 3, which is the season 2020/2021, where the majority of matches were played behind closed doors due to the COVID-19 restrictions. This season in approximately 60.1% of games the away team managed not to lose the match. This was most likely due to the partial loss of home advantage for the home team as there were no or the bare minimum of fans allowed in the stadium. Compared to the overall ratio of results, the difference was around 3.5%. On the other hand, in season 5 home team managed to win 45.7% of games, which was the most from all 5 seasons, and it was around 2.3% more than the overall average.

For the complete overview of all variables, we can look at Table 3.9, which is the summary table of all numerical variables that belong to the group of traditional variables, and Table 3.10, where are all xG variables that were created. The names of the variables are explained in the Appendix A. If we look at the odds we gathered, we see that the maximal closing odds, which are supposed to

Season	Total Matches	Home Win(%)	Draw/Away Win(%)
1	1826	44.7	55.3
2	1725	44.1	55.9
3	1826	39.9	60.1
4	1826	42.8	57.2
5	1826	45.7	54.3
Overall	9029	43.4	56.6

Table 3.8: Seasonal and Overall Percentages of Match Outcomes

be higher than odds from one betting site, are smaller. However, this is caused by the fact that for the first season (2018/2019), Football-Data.co.uk[8] did not have this variable available, so the variables MaxCH and MaxCDA are set to 0, which affects the summary statistics. For the following four seasons, the mean odds from the betting site bet365[9] remain nearly the same, but the mean of maximal closing odds is approximately 3.09 for the home team (MaxCH) and 2.03 for the match to result in a draw or an away team win (MaxCDA). Thus, we get higher odds using the maximal closing odds, which is logical as the bettor is supposed to gain an advantage by comparing the betting market and using the highest odds found possible.

[8]www.football-data.co.uk/
[9]https://www.bet365.com/

Variable	Min	1st Qu.	Mean	3rd Qu.	Max
Season	1.000	2.000	3.011	4.000	5.000
GameWeek	1.00	10.00	19.01	28.00	38.00
HomeELO	0.5924	0.8564	1.0000	1.1129	1.6537
AwayELO	0.5924	0.8564	0.9999	1.1129	1.6537
HomePreMatchPosition	0.00	5.00	10.13	15.00	20.00
AwayPreMatchPosition	0.00	5.00	9.971	15.00	20.00
FTR	0.0000	0.0000	0.4344	1.0000	1.0000
B365H	1.030	1.660	2.808	3.100	23.000
B365DA	1.000	1.370	1.976	2.200	12.980
MaxCH	0.000	1.330	2.472	3.120	35.000
MaxCDA	0.000	1.130	1.617	2.020	17.050
HTAS	0.0000	0.3529	0.4837	0.6000	2.0339
ATAS	0.0000	0.3563	0.4897	0.6061	2.2785
HTDW	0.0000	0.3846	0.4885	0.5970	2.0000
ATDW	0.0000	0.3796	0.4833	0.5951	1.8519
AvgHTP	0.0000	0.9459	1.3256	1.7143	3.0000
AvgATP	0.0000	0.9655	1.3443	1.7273	3.0000
AvgHTGD	-8.0000	-0.53846	-0.01270	0.47059	5.0000
AvgATGD	-5.0000	-0.50000	0.01839	0.50000	8.0000
HomeForm	0.0000	0.2674	0.4334	0.6007	1.2577
AwayForm	0.0000	0.2765	0.4475	0.6168	1.5142
HomeTeamH2HRatio	0.0000	0.0000	0.2877	0.4000	1.0000
AwayTeamH2HRatio	0.0000	0.0000	0.2990	0.4000	1.0000
AvgHTSOT	0.000	3.529	4.321	5.091	15.000
AvgATSOT	0.000	3.559	4.367	5.125	15.000
AvgHTC	0.000	4.125	4.754	5.478	16.000
AvgHTCAgainst	0.000	4.121	4.802	5.600	15.000
AvgATC	0.000	4.161	4.808	5.500	15.000
AvgATCAgainst	0.000	4.080	4.752	5.545	16.000
AvgHTF	0.000	10.88	12.04	13.64	27.00
AvgATF	0.000	10.88	12.02	13.62	24.00
AvgHTFAgainst	0.000	11.00	12.03	13.62	27.00
AvgATFAgainst	0.000	11.00	12.03	13.62	27.00
HomeWinRatioSoFar	0.0000	0.2222	0.4005	0.5625	1.0000
AwayNotLosingRatioSoFar	0.0000	0.3889	0.5464	0.7273	1.0000

Table 3.9: Summary of Traditional Variables - Explanation of Shortcuts in Appendix A.1

Variable	Min	1st Qu.	Mean	3rd Qu.	Max
HTxGASFooty	0.0000	0.4357	0.4835	0.5437	0.9622
ATxGASFooty	0.0000	0.4393	0.4881	0.5476	1.1731
HTxGASUnder	0.0000	0.3830	0.4829	0.5803	1.5463
ATxGASUnder	0.0000	0.3865	0.4894	0.5867	1.5408
HTxGDWFooty	0.0000	0.4760	0.5297	0.6002	1.2586
ATxGDWFooty	0.0000	0.4725	0.5254	0.5964	1.4099
HTxGDWUnder	0.0000	0.4068	0.4882	0.5780	1.5408
ATxGDWUnder	0.0000	0.4027	0.4830	0.5743	1.5463
HTAEFooty	0.0000	0.7139	0.8828	1.0558	3.2609
ATAEFooty	0.0000	0.7161	0.8854	1.0562	3.5714
HTAEUnder	0.0000	0.8483	0.9732	1.1111	11.1111
ATAEUnder	0.0000	0.8475	0.9708	1.1111	11.1111
HTDEFooty	0.0000	0.7435	0.8925	1.0584	3.5714
ATDEFooty	0.0000	0.7382	0.8915	1.0619	3.2609
HTDEUnder	0.0000	0.8522	0.9726	1.1168	11.1111
ATDEUnder	0.0000	0.8511	0.9757	1.1176	11.1111
HomexGShotFooty	0.0000	0.1180	0.1234	0.1313	0.2913
AwayxGShotFooty	0.0000	0.1178	0.1232	0.1312	0.2525
HomexGShotUnder	0.00000	0.09706	0.11196	0.12879	0.44500
AwayxGShotUnder	0.00000	0.09714	0.11223	0.12894	0.49000
HomeFootyForm	-3.35287	-0.44845	-0.06972	0.31285	2.34001
AwayFootyForm	-3.42035	-0.34660	0.02927	0.41321	2.23223
HomeUnderForm	-6.85421	-0.66067	-0.08814	0.49498	4.17353
AwayUnderForm	-6.00113	-0.55733	0.02734	0.61177	4.26829
HomeFootyPos	0.00	5.00	10.18	15.00	20.00
AwayFootyPos	0.00	5.00	9.922	15.00	20.00
HomeUnderPos	0.00	5.00	10.15	15.00	20.00
AwayUnderPos	0.00	5.00	9.949	15.00	20.00

Table 3.10: Summary of xG Variables - Explanation of Shortcuts in Appendix A.2

Chapter 4

Methodology

4.1 Logistic Regression

The goal of this study is to test the predictions of football matches made with logistic regression using different inputs (combinations of predictive variables). As a result of defining the outcome variable FTR as binary (either a home team winning or an away team not losing), and the existing literature, the binary logistic regression was used to gather predictions that were converted to odds and then used for developing a betting strategy.

The logistic regression is a limited dependent variable model that is used to estimate the effect of independent variables on a restricted dependent variable, which is FTR in this case. It is often used to strictly limit the estimated probabilities between 0 and 1 (Wooldridge (2012)). Thus, we define the binary response model as follows:

$$P(y_i = 1 \mid \mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = G(\beta_0 + \mathbf{x}^\top \beta),$$

where $\mathbf{x}$ is a matrix of all independent variables and G is the function taking values strictly between zero and one: $0 < G(z) < 1$. In our case, for the logistic regression, the function $G(z)$ refers to:

$$G(z) = \frac{exp(z)}{1 + exp(z)}$$

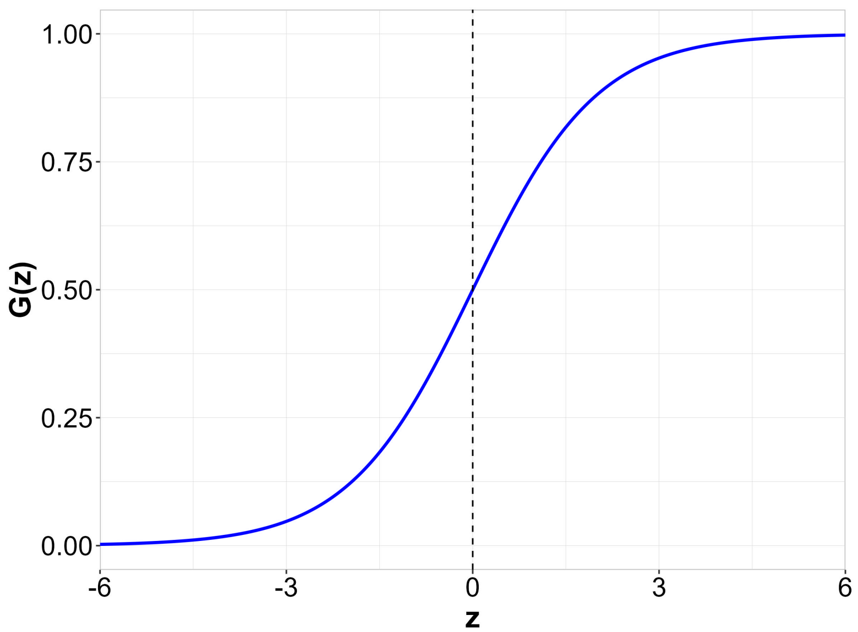

Figure 4.1: Graph of the Logistic Function $G(z) = \frac{exp(z)}{1+exp(z)}$

In Figure 4.1 above we see the graph of the logistic function $G(z)$. We can see that it is strictly between 0 and 1 and $G(z) \rightarrow 0$ as $z \rightarrow -\infty$ and $G(z) \rightarrow 1$ as $z \rightarrow +\infty$. The logistic function increases most quickly at $z = 0$ (Wooldridge (2012)).

The logistic model can be derived from an underlying latent variable model. We consider an unobserved (latent) variable y^*, which is linearly influenced by explanatory variable x:

$$y^* = \beta_0 + \beta_1 x + \epsilon,$$

we observe only y, which is a binary variable such that:

$$y = 1[y^* > 0],$$

where $1[y^* > 0]$ is called indicator function, which takes the value of 1 if something happens, 0 otherwise. In this case, if $y^* > 0$, then the function will have the value 1, otherwise it will have the value 0.

Then:

$$\begin{aligned}P(y=1 \mid x) &= P(y^* > 0 \mid x) = P(\epsilon > -(\beta_0 + \beta_1 x) \mid x)\\ &= 1 - G[-(\beta_0 + \beta_1 x)]\\ &= G(\beta_0 + \beta_1 x)\end{aligned}$$

Due to the nonlinear nature of the logistic function, we have to use the Maximum Likelihood Estimation (MLE). To obtain the MLE, we first need a density of y_i given x_i:

$$f(y \mid x_i; \beta) = [G(x_i\beta)]^y [1 - G(x_i\beta)]^{1-y}, \quad \text{where } y = 0 \text{ or } y = 1.$$

Lastly, to obtain the MLE of β, we need to maximize the following log-likelihood:

$$L(\beta) = \sum_{i=1}^{n} y_i \log[G(x_i\beta)] + (1 - y_i) \log[1 - G(x_i\beta)],$$

This function is obtained by applying the logarithm to the likelihood function for a single observation and using the property of logarithm $log(a^b) = blog(a)$. Then the log-likelihood for all n observations is the sum of all log-likelihoods for observations indexed by i starting from 1 to n. From the theory of conditional MLE, under very general conditions, the MLE is consistent, asymptotically normal, and asymptotically efficient.

The difficult aspect of the logistic regression model is presenting and interpreting the results. As we are usually interested in the effects of independent variables x_j on the response probability $P(y = 1 \mid \mathbf{x})$, but this is complicated by the nonlinear nature of the logistic function. To find the partial effect of roughly continuous variables on the response model, we have to use the partial derivative:

$$\frac{\partial P(y = 1 \mid \mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\beta)\beta_j$$

Thus, the β_j now represents the effect of a jth independent variable on the response probability $P(y = 1 \mid x)$, but their magnitudes do not have direct interpretation since the effect of x on $P(y = 1 \mid x)$ is not constant as it depends on all other variables $\mathbf{x}$. This implies that by using the partial derivative, we can directly address only the direction of the effect, but not the magnitude. This means that the positive/negative sign of $\hat{\beta}_j$ indicated that an increase in the jth independent variable increases/decreases the response probability $P(y = 1 \mid x)$, but still, we do not know by how much.

However, as we are interested in betting and creating the most accurate model, we do not necessarily need the magnitudes, we are more interested in the probability of outcome and the sign of coefficients to ensure that the direction of the effect of all variables is as expected.

4.2 Assumptions

According to Schreiber-Gregory & Bader (2018) the logistic regression assumptions differ from the linear regression assumptions. Firstly, logistic regression does not require a linear relationship between the dependent and independent variables, the error terms (residuals) do not need to be normally distributed and lastly, homoskedasticity is not required. However, it still shares some assumptions, with some additional assumptions that have to be verified.

The first assumption is the structure of the outcome variable, which has to be binary for binary logistic regression to work. This assumption is met based on the creation of the outcome variable FTR in Section 3.2. The second assumption is observation independence, which states that the observations should be independent of each other, which is also met by the nature of our data and the fact that every football match is an independent event. The third assumption is about the absence of multicollinearity, which means that the independent variables should not be too highly correlated with each other. This assumption will be tested in Section 4.4 along with the fourth assumption about the linearity of independent variables and log odds. Unlike the linearity of dependent and independent variables, logistic regression requires that the independent variables are linearly related to the log odds. Lastly, the assumption of a large sample size has to be met. The authors of this study follow a general guideline that you need a minimum of 10 cases with the least frequent outcome for each independent variable in your model. However, Bujang *et al.* (2018) in their paper about sample size guidelines for logistic regression, recommend that the minimum sample size should be 500 observations and also proposed a formula $n = 100 + 50i$, where i refers to the number of independent variables in the final model. As my dataset contains 9029 matches, this assumption is satisfied for my models. If all assumptions are met, the models introduced in the next section are consistent, asymptotically normal, and asymptotically efficient (Wooldridge (2012)).

4.3 Rolling Training/Testing Set

We will split our dataset into training and testing sets as it is essential to properly examine the model's ability to generalize to new data as confirmed in many studies (Rácz *et al.* (2021), Singh *et al.* (2021), Nguyen *et al.* (2021))

Our dataset contains 5 seasons across 5 different leagues. As the commonly used split ratios are 70/30 and 80/20 for the training set, we will test the performance of the models in the last 2 seasons. So for the start of season 4, the models are trained on data from season 1 to season 3. Our main goal is to create a profitable betting strategy, we want to train the model as much as possible to achieve the best results. Also as the football matches are mostly played during the weekend, we will bet on the results every game week. This allows us to shift the training set by one week after each game week is played. The technique is explained in the Table 4.1.

Season	**Game Week**	**Training Set**
4	1	Seasons 1-3
4	2	Seasons 1-3 + Season 4 GW 1
4	3	Seasons 1-3 + Season 4 GW 1-2
...	...	...
5	1	Seasons 1-4
5	2	Seasons 1-4 + Season 5 GW 1
5	3	Seasons 1-4 + Season 5 GW 1-2
...	...	...

Table 4.1: Illustration of Training and Testing Set

As illustrated above, after the first game week of season 4, we will move the game week 1 data into the training set, so that the model can learn on it. This is then replicated for the rest of the season. The same procedure is applied to season 5 analogously to the season 4 example that was just explained. In the following Section 4.4, the results of the regression are demonstrated on the original training dataset (Season 1 - Season 3) as the values of coefficients are not changing significantly. For the development of coefficients over time, you can see the graphs in Appendix B.

4.4 Modelling

As existing literature covers predicting outcomes of football matches using traditional variables, we want to compare the performance of models that include xG and those that do not. As we have two sources of xG statistics, 5 models will be in the following subsections.

4.4.1 Traditional Model

This model will use variables based on the literature. Most of the variables that were created in Section 3.2. As van Wijk (2021) conducted the most detailed research on football predictions, we created a model that is based on the variables he used. However, there are several changes and limitations. First of all, he uses a historical H2H ratio (record of two teams against each other) that includes all matches in the history between two given teams and the ELO that both teams have before each match. These two variables are not publicly available in a format that could be used, so the variables in my model are adjusted. Firstly, my H2H ratio is calculated from the last 5 matches as explained in Section 3.2, and the ELO variable has the value that was assigned to each team at the beginning of each season.

Additionally, the presence of many variables increases the chances of violating the multicollinearity assumption in Section 4.2, which would make the model's results unreliable. Therefore, the following correlation matrix is provided. However, the variables are calculated for both home and away teams, so for easier visualization, there are only the variables for the home team included, as the correlation is approximately the same for both home and away variables. In Figure 4.2 we can see that many variables are highly correlated, so we will use only the most important variables. The variables $AvgHTP$ and $AvgHTGD$ are strongly correlated with more than one other variable, so we will exclude them from our model. Additionally, $AvgHTSOT$ are highly correlated with $HTAS$, and as the attacking strength has a more significant effect on the outcome of the match than an average number of shots on target, we will exclude $AvgHTSOT$ from the model. We are left with few other variables, from which only $HomeELO$, $HTAS$, $HTDW$, $AvgHTC$, and $AvgHTCAgainst$ remain significant during the regression when put together in one model as the ELO variable seems to be the most strong predictor by far. Thus, for this model, we will use only these (and also variables for the away team) in the future.

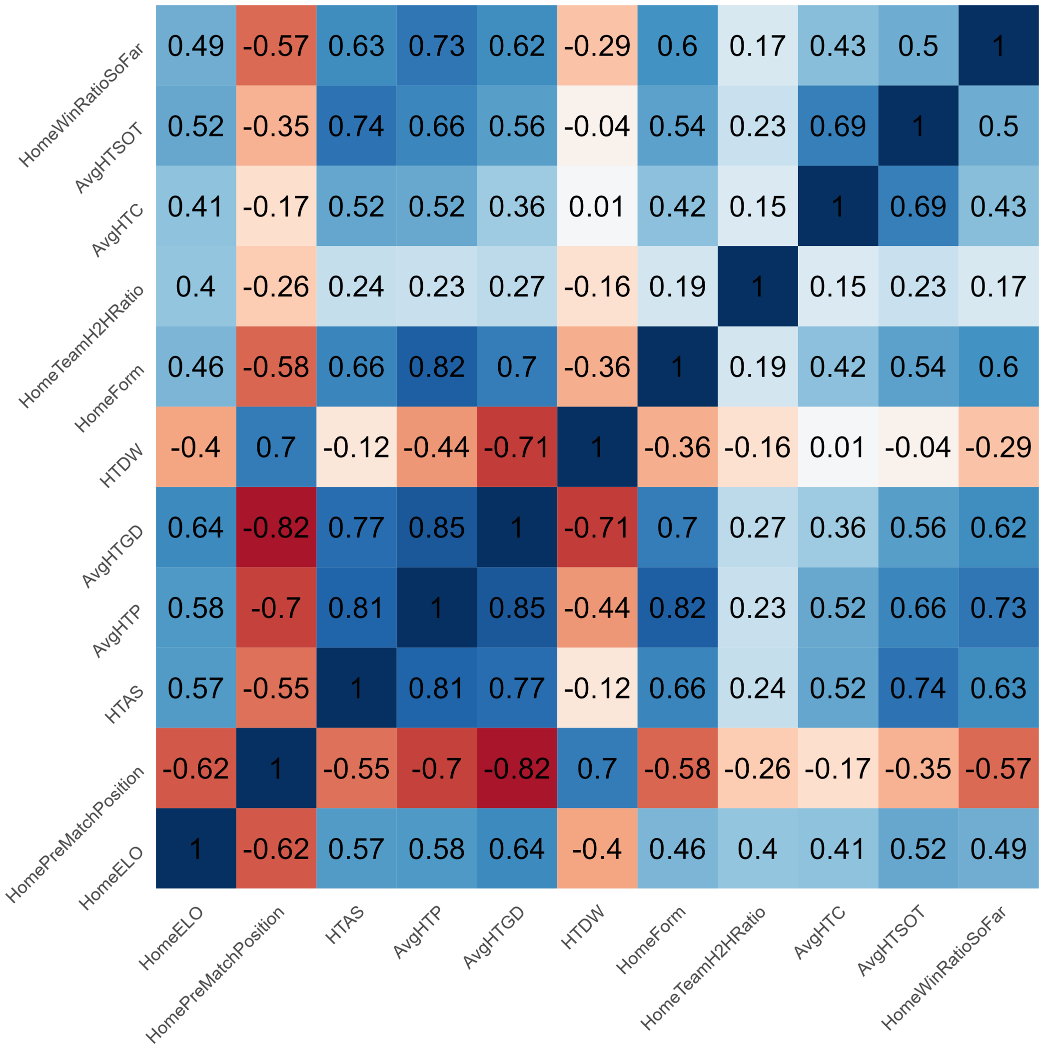

Figure 4.2: Correlation between Traditional Variables

So the Traditional Model will be defined as follows:

$$\begin{aligned}\log\left(\frac{P(FTR_i=1)}{1-P(FTR_i=1)}\right) = \beta_0 + \beta_1 HomeELO_i + \beta_2 AwayELO_i + \\ \beta_3 HTAS_i + \beta_4 ATAS_i + \beta_5 HTDW_i + \\ \beta_6 ATDW_i + \beta_7 AvgHTC_i + \beta_8 AvgATC_i + \\ \beta_9 AvgHTCAgainst_i + \beta_{10} AvgATCAgainst_i\end{aligned} \tag{4.1}$$

We expect a positive sign for $HomeELO$ as stronger teams have higher values of ELO. Moreover, $HTAS$, $AvgHTC$ should have a positive sign as well, because attacking strength is calculated in a way that higher values indicate the stronger offensive ability of a team as explained in Section 3.2 and more corners per game indicate more chances for a team to score and win a match. Therefore, *AvgHTCAgainst* is expected to have a negative sign as corners against the team are an opportunity for the opposing team to score. We expect $HTDW$ to have a negative sign as lower values indicate strong defending from a team. For the away team variables, we expect the opposite signs to the home team variables.

As the model still contains a lot of variables and there were some potential correlations, we will use another approach used for identifying multicollinearity in the model called the Variance Inflation Factor (VIF). The formula for VIF is as follows:

$$VIF_i = \frac{1}{1 - R_i^2},$$

where R_i^2 = unadjusted coefficient of determination for regression of the ith independent variable on the remaining ones. According to Craney & Surles (2002) in their study, the value 10 is critical for keeping the variable in the model. However Daoud (2017) states that for $VIF > 5$ the correlation is strong and the model could be struggling with multicollinearity issues. Thus, we will use the VIF after running the regression on the data from season 1 to season 3 (because seasons 4 and 5 are already included in our testing set) to make sure there are no strongly correlated variables before making predictions.

Variable	VIF
HomeELO	1.88
AwayELO	1.86
HTAS	1.75
ATAS	1.67
HTDW	1.43
ATDW	1.43
AvgHTC	1.77
AvgATC	1.71
AvgHTCAgainst	1.64
AvgATCAgainst	1.63

Table 4.2: VIF Traditional Model

The maximum value of VIF is for HomeELO (1.88), which suggests that there is no correlation between two variables that would possibly hurt the reliability of this model's results.

4.4.2 Understat xG Model

The second model is created only from variables created using the xG values obtained from Understat. As the variables for the xG models were created while trying to capture new information with each created variable, we will put all variables into the xG model as there is no expected multicollinearity issue.

$$\begin{aligned}\log\left(\frac{P(FTR_i=1)}{1-P(FTR_i=1)}\right) = \beta_0 + \beta_1 HomeUnderPos_i + \\ \beta_2 AwayUnderPos_i + \beta_3 HTxGASUnder_i + \\ \beta_4 ATxGASUnder_i + \beta_5 HomexGShotUnder_i + \\ \beta_6 AwayxGShotUnder_i + \beta_7 HTxGDWUnder_i + \\ \beta_8 ATxGDWUnder_i + \beta_9 HomeFormUnder_i + \\ \beta_{10} AwayFormUnder_i + \beta_{11} HTAEUnder_i + \\ \beta_{12} ATAEUnder_i + \beta_{13} HTDEUnder_i + \\ \beta_{14} ATDEUnder_i\end{aligned} \tag{4.2}$$

Similarly to the traditional model, we expect the same signs for the attacking strength and defensive weakness. Moreover, the sign of $HTAEUnder$ should be positive as higher values indicate better efficiency in scoring goals according to Section 3.3.3. $HTDE$ is the opposite as lower values indicate better defending efficiency, thus negative sign is expected. The better the form, the higher the value of $HomeUnderForm$, so a positive sign is anticipated. For the position in the league based on xG $HomeUnderPos$, we expect a negative sign, because teams placed higher in the league are expected to win their games. The sign of $HomexGShotUnder$ is debatable as we could argue that great teams create many chances and have a lot of shots, so the expected sign would be negative. However, it could also be positive as lower values would indicate that the team is effective in converting their shots into a goal. For the away team variables, the opposite signs are assumed.

To make sure the variables are not strongly correlated, we calculate the variance inflation factor to detect multicollinearity in our model. In Table 4.3, we can see the higher value of VIF is for $HTxGASUnder$ (3.74), which suggests a moderate correlation. However, values below 5 should not indicate any potential problems with multicollinearity as explained in 4.4.1.

Variable	VIF
HomeUnderPos	1.89
AwayUnderPos	1.96
HTxGASUnder	3.74
ATxGASUnder	3.60
HomexGShotUnder	3.28
AwayxGShotUnder	3.09
HTxGDWUnder	2.48
ATxGDWUnder	2.47
HomeUnderForm	2.12
AwayUnderForm	2.11
HTAEUnder	1.24
ATAEUnder	1.23
HTDEUnder	1.20
ATDEUnder	1.24

Table 4.3: VIF xG Understat Model

4.4.3 FootyStats xG Model

The FootyStats model defined in equation 4.3 is created analogously to the Understat one. The highest VIF value is for the xG per shot variables, particularly for the home team variable (4.62), which is still below the critical threshold, thus we keep it in the model.

$$\begin{aligned}\log\left(\frac{P(FTR_i=1)}{1-P(FTR_i=1)}\right) = \beta_0 &+ \beta_1 HomeFootyPos_i + \\ &\beta_2 AwayFootyPos_i + \beta_3 HTxGASFooty_i + \\ &\beta_4 ATxGASFooty_i + \beta_5 HomexGShotFooty_i + \\ &\beta_6 AwayxGShotFooty_i + \beta_7 HTxGDWFooty_i + \\ &\beta_8 ATxGDWFooty_i + \beta_9 HomeFootyForm_i + \\ &\beta_{10} AwayFootyForm_i + \beta_{11} HTAEFooty_i + \\ &\beta_{12} ATAEFooty_i + \beta_{13} HTDEFooty_i + \\ &\beta_{14} ATDEFooty_i\end{aligned} \tag{4.3}$$

The signs are expected to be the same as in the Understat xG model, as the variables are calculated in the same manner, only with the different source of xG values. Table 4.4 shows that for other variables apart from xG per shot, the VIF is smaller, so we have no issues with multicollinearity.

Variable	VIF
HTxGASFooty	3.22
ATxGASFooty	3.29
HTxGDWFooty	3.51
ATxGDWFooty	3.47
HTAEFooty	1.39
ATAEFooty	1.35
HTDEFooty	1.29
ATDEFooty	1.29
HomeFootyForm	2.24
AwayFootyForm	2.21
HomeFootyPos	1.95
AwayFootyPos	1.96
HomexGShotFooty	4.62
AwayxGShotFooty	4.54

Table 4.4: VIF xG FootyStats Model

Overview of the 3 Basic Models

As we now want to create the last two models that will include the most important variables from both the traditional and xG models, we have to look at the summary tables of these models. The summary tables are from regressing all 3 models on the data from Season 1 to Season 3, as we then adjust the training set after each game week, which was explained in Section 4.3. The development of coefficients in time is plotted and can be seen in Appendix B.

Predictors	Estimate	Std. Error	z value	p value	Significance
HomeELO	2.076	0.204	10.186	< 2e-16	***
AwayELO	-1.390	0.209	-6.662	2.70e-11	***
HTAS	0.348	0.187	1.865	0.06219	.
ATAS	-0.490	0.184	-2.657	0.00789	**
HTDW	-0.556	0.188	-2.953	0.00315	**
ATDW	0.561	0.191	2.941	0.00327	**
AvgHTC	0.068	0.028	2.410	0.01594	*
AvgATC	-0.065	0.027	-2.381	0.01728	*
AvgHTCAgainst	-0.077	0.026	-2.928	0.00341	**
AvgATCAgainst	0.103	0.026	3.912	9.14e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.5: Coefficients Traditional Model

Predictors	Estimate	Std. Error	z value	p value	Significance
HomeUnderPos	-0.020	0.007	-2.903	0.004	**
AwayUnderPos	0.013	0.007	1.899	0.058	.
HTxGASUnder	3.157	0.329	9.592	< 2e-16	***
ATxGASUnder	-2.453	0.329	-7.453	9.16e-14	***
HomexGShotUnder	-8.907	1.604	-5.552	2.82e-08	***
AwayxGShotUnder	4.751	1.569	3.028	0.002	**
HTxGDWUnder	-1.250	0.303	-4.123	3.73e-05	***
ATxGDWUnder	1.551	0.303	5.115	3.14e-07	***
HomeUnderForm	0.049	0.048	1.013	0.311	
AwayUnderForm	-0.005	0.049	-0.093	0.926	
HTAEUnder	0.206	0.092	2.238	0.025	*
ATAEUnder	-0.197	0.094	-2.107	0.035	*
HTDEUnder	-0.159	0.087	-1.835	0.067	.
ATDEUnder	0.129	0.091	1.423	0.155	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.6: Coefficients Understat Model

Predictors	Estimate	Std. Error	z value	p value	Significance
HomeFootyPos	-0.018	0.007	-2.479	0.013	*
AwayFootyPos	0.009	0.007	1.219	0.223	
HTxGASFooty	3.400	0.436	7.796	6.38e-15	***
ATxGASFooty	-2.703	0.442	-6.117	9.54e-10	***
HomexGShotFooty	-3.633	2.242	-1.621	0.105	
AwayxGShotFooty	-0.777	2.205	-0.352	0.725	
HTxGDWFooty	-1.843	0.427	-4.318	1.57e-05	***
ATxGDWFooty	2.518	0.426	5.917	3.29e-09	***
HomeFootyForm	0.029	0.077	0.380	0.704	
AwayFootyForm	-0.051	0.076	-0.671	0.502	
HTAEFooty	0.364	0.104	3.492	0.00048	***
ATAEFooty	-0.287	0.104	-2.768	0.0056	**
HTDEFooty	-0.433	0.105	-4.134	3.57e-05	***
ATDEFooty	0.215	0.103	2.092	0.036	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.7: Coefficients FootyStats Model

The tables above represent the results of our logistic regression models. In Table 4.5, we can see that all variables in the Traditional model are significant at least at 90% confidence level. The variables *HomeELO* and *AwayELO* have the smallest p-values of all variables, so they are strong predictors as expected. Additionally, the signs of all coefficients are as expected. Table 4.6 represents the coefficients in the Understat model and we see that the attacking strength and defensive weakness have the smallest p-values. Additionally, the xG per shot variables are highly significant. On the other hand, the form variables are extremely insignificant and the defensive efficiency is only significant at 90% confidence level for the home team. The signs are as expected, the only question we had was about the sign of the $HomexGShotUnder$ and $AwayxGShotUnder$ variables, which indicate that if the team has higher xG per shot, they are less likely to win/not lose a match. From Table 4.7, we conclude that similar things apply to the FootyStats model, where attacking strength and defensive weakness have the smallest p-values. Additionally, the form variables are also insignificant. For both FootyStats and Understat models, the home team position is significant. In contrast, for Understat the away position is only significant at a 90% confidence level and for FootyStats it is not significant at any typically used level. However, for FootyStats the attacking and defensive efficiency are statistically significant, which is not the case for $HomexGShotFooty$ and $AwayxGShotFooty$, unlike the Understat model.

The last important thing we need to check is the assumption of the linear relationship between log odds and the independent variables from Section 4.2. This assumption can be tested with two different approaches: Box-Tidwell-Test or by plotting the relationship between the log odds and the independent variables (Kassambara (2018), Leung (2021)). Thus, it is tested by plotting scatterplots of independent variables against log odds.

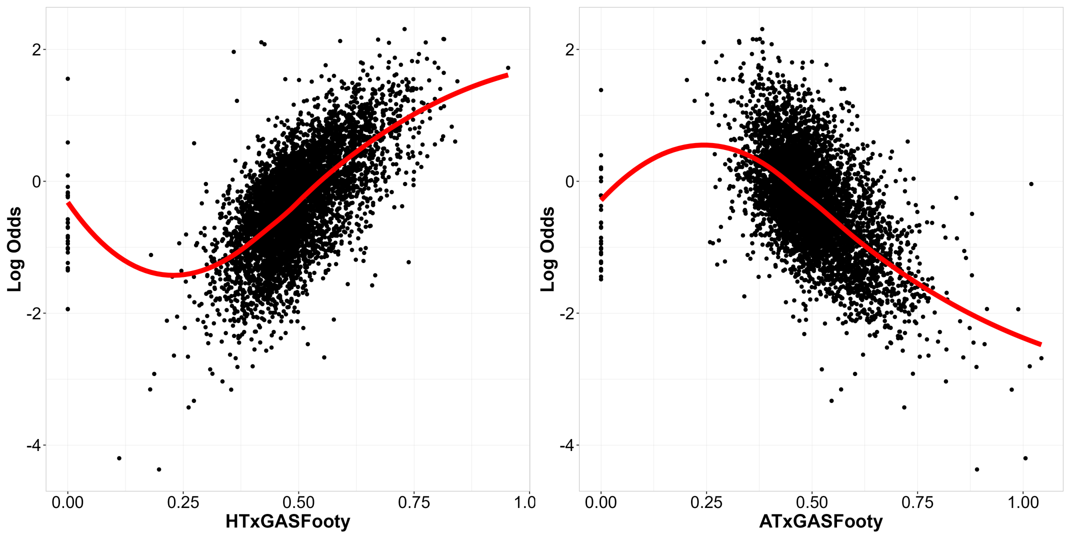

Figure 4.3: Log Odds - Independent Variables Plots

As we can see in Figure 4.3 where is plotted the attacking strength of the home team and then the attacking strength of the away team using the xG values against the log odds using the lowess curve, the 0s (which are in the data due to our creation of pre-match variables, where for the first matches, most of them are set to 0) are skewing the linear relationship, so the assumption is not exactly satisfied. However, the accuracies of models using a dataset with zeros and then a dataset where the matches with variables set to 0 are removed, are approximately the same. So if we ignore the zeros as outliers, the assumption is satisfied, as we can see the lowess curve is roughly linear for the rest of the dataset. The assumption was tested for all continuous variables as it does not make sense for categorical variables. For example, for variables $HomeELO$ and $AwayELO$, which are not set to 0 for the first matches of each season, Figure 4.4 demonstrates the linear relationship for both.

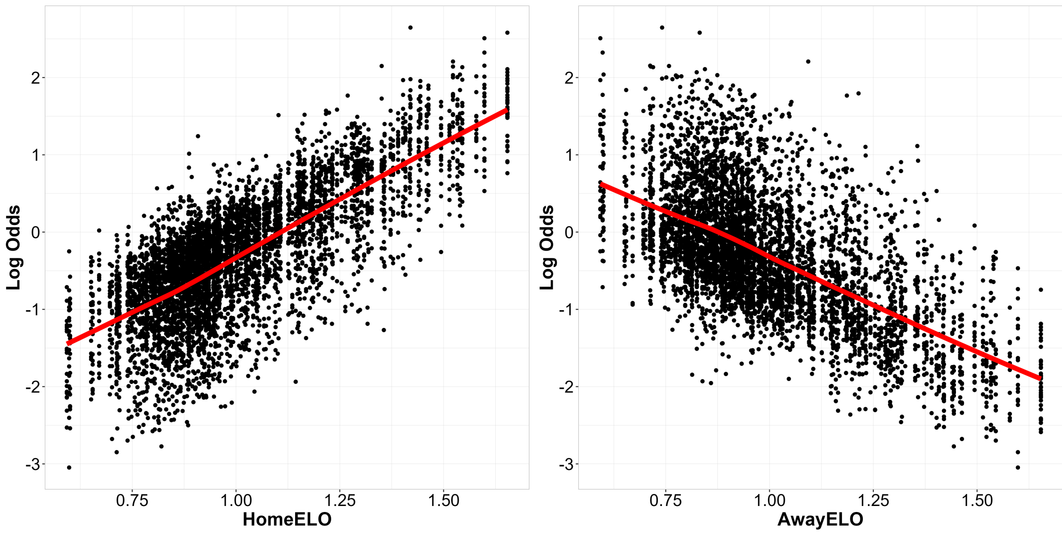

Figure 4.4: Log Odds - ELO Variables Plots

4.4.4 Combined Models

Now, we create two models that will be a combination of traditional variables and xG variables to capture the predictive power of these variables when put together. We will create a model that will be a combination of traditional variables and Understat xG variables and then the traditional variables combined with FootyStats xG variables. Firstly, we have to decide what variables to use, it would make sense to include all significant variables from both models into one model. However, this would not work as attacking strength and defensive weakness is highly correlated when calculated from actual goals and expected goals, which is confirmed in Table 4.8. Thus, we have to use some domain knowledge about football to estimate which predictors will work best together. The variable that we will use in both models is the ELO variable as it captures the team's relative strength based on historical performances and its predictive power is confirmed in Table 4.5. As the significance of attacking strength and defensive weakness in the traditional model is not that high compared to the ELO variable, we will use the attacking strength and defensive weakness based on xG in our combined models.

Variables	**Correlation Coefficient**
HTAS vs. HTxGASUnder	0.835
HTDW vs. HTxGDWUnder	0.802

Table 4.8: Correlation Coefficients between Attacking Strength and Defensive Weakness based on xG and Actual Goals

Combined Understat Model

Thus, for the model with traditional variables and variables from Understat xG, we will use as explained above the ELO variable, the attacking strength and the defensive weakness based on xG from Understat. Additionally, we can include the xG per shot variables as they were significant in the Understat model along with the attacking efficiency of both teams. We will also include the average number of corners for and against the teams. Therefore, the combined model is defined as follows:

$$
\begin{aligned}
\log\left(\frac{P(FTR_i = 1)}{1 - P(FTR_i = 1)}\right) = \beta_0 &+ \beta_1 HomeELO_i + \beta_2 AwayELO_i \\
&\beta_3 HTxGASUnder_i + \beta_4 ATxGASUnder_i \\
&\beta_5 HomexGShotUnder_i + \beta_6 AwayxGShotUnder_i \\
&\beta_7 HTxGDWUnder_i + \beta_8 ATxGDWUnder_i \\
&\beta_9 HTAEUnder_i + \beta_{10} ATAEUnder_i \\
&\beta_{11} AvgHTC_i + \beta_{12} AvgATC_i \\
&\beta_{13} AvgHTCAgainst_i + \beta_{14} AvgATCAgainst_i
\end{aligned}
\tag{4.4}
$$

The value of the variance inflation factor represented by Table 4.9 for *HTxGA-SUnder* is 4.95 and 4.59 for the $ATxGASUnder$, which is caused mainly by the correlation with the xG per shot variable. However, the value is still strictly below the critical threshold of 5, and mainly the attacking strength is essential for predicting football matches, so we will keep the variable in the model.

Predictor	**VIF**
HomeELO	1.90
AwayELO	1.90
HTxGASUnder	4.95
ATxGASUnder	4.59
HTxGDWUnder	2.01
ATxGDWUnder	2.05
HomexGShotUnder	3.75
AwayxGShotUnder	3.49
HTAEUnder	1.22
ATAEUnder	1.25
AvgHTC	2.30
AvgATC	2.15
AvgHTCAgainst	2.13
AvgATCAgainst	2.13

Table 4.9: VIF Combined Understat Model

Combined FootyStats Model

Analogously, the model with combined traditional variables and FootyStats xG variables will be created. From the Traditional model, the ELO and corners for and against the team will be used in the combined model again. From the FootyStats model, we will use the attacking strength and defensive weakness alongside both attacking and defensive efficiency as all these variables were significant in the FootyStats model.

$$\begin{aligned}\log\left(\frac{P(FTR_i=1)}{1-P(FTR_i=1)}\right) = \beta_0 + \beta_1 HomeELO_i + \beta_2 AwayELO_i \\ \beta_3 HTxGASFooty_i + \beta_4 ATxGASFooty_i \\ \beta_5 HTxGDWFooty_i + \beta_6 ATxGDWFooty_i \\ \beta_7 HTAEFooty_i + \beta_8 ATAEFooty_i \\ \beta_9 HTDEFooty_i + \beta_{10} ATDEFooty_i \\ \beta_{11} AvgHTC_i + \beta_{12} AvgATC_i \\ \beta_{13} AvgHTCAgainst_i + \beta_{14} AvgATCAgainst_i\end{aligned} \tag{4.5}$$

Similarly, to the Combined Understat Model, we can see in Table 4.10 that the highest values of VIF are for the attacking strength variables. Nevertheless, the value is 4.27, sufficiently low for keeping it in the model.

Predictor	**VIF**
HomeELO	1.95
AwayELO	1.91
HTxGASFooty	4.27
ATxGASFooty	4.18
HTxGDWFooty	3.41
ATxGDWFooty	3.47
HTAEFooty	1.45
ATAEFooty	1.42
HTDEFooty	1.37
ATDEFooty	1.39
AvgHTC	2.81
AvgATC	2.70
AvgHTCAgainst	2.58
AvgATCAgainst	2.66

Table 4.10: VIF Combined FootyStats Model

4.4.5 Overview of the Combined Models

Lastly, we look at the coefficients and their significance in our two combined models represented by Table 4.11 and Table 4.12. We can see that in both models, the variable ELO has the smallest p-value by far. Attacking strength and defensive weakness are significant in both models, but the rest of the variables seem to not have a significant effect on the outcome variable. This could be explained by the fact that ELO represents the relative strength of a team based on historical performances and when combined with the attacking strength and defensive weakness so far in the season, these variables capture the performance of a team well enough. However, for the Combined Understat model, we can see that the $HomexGShotUnder$ and $AvgATCAgainst$ are significant. For the Combined FootyStats model, the only other significant variable is $HTDEFooty$. As the variables are significant, we will keep them in the model along with the same variables but for the opposing team. Thus, the models are defined as in the equations 4.6 and 4.7. These models will be compared with the others in Chapter 6. We do not have to control for multicollinearity using the Variance Inflation Factor as we did not introduce new variables in the model, on the contrary, we removed some.

Predictors	Estimate	Std. Error	z value	p value	Significance
(Intercept)	-0.988	0.328	-3.013	0.0026	**
HomeELO	1.950	0.204	9.549	< 2e-16	***
AwayELO	-1.287	0.210	-6.128	8.89e-10	***
HTxGASUnder	1.431	0.384	3.729	0.00019	***
ATxGASUnder	-1.018	0.376	-2.706	0.0068	**
HTxGDWUnder	-0.940	0.272	-3.462	0.00054	***
ATxGDWUnder	0.978	0.277	3.535	0.00041	***
HomexGShotUnder	-4.772	1.733	-2.753	0.0059	**
AwayxGShotUnder	0.033	1.697	0.019	0.985	
AvgHTC	0.028	0.032	0.890	0.374	
AvgATC	-0.011	0.030	-0.363	0.716	
AvgHTCAgainst	-0.026	0.030	-0.878	0.380	
AvgATCAgainst	0.079	0.030	2.632	0.0085	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.11: Coefficients Combined Understat Model

Predictors	Estimate	Std. Error	z value	p value	Significance
(Intercept)	-1.07	0.334	-3.20	0.0014	**
HomeELO	1.94	0.208	9.32	< 2e-16	***
AwayELO	-1.21	0.212	-5.72	1.04e-08	***
HTxGASFooty	1.28	0.505	2.53	0.0115	*
ATxGASFooty	-1.76	0.503	-3.49	0.00048	***
HTxGDWFooty	-1.33	0.412	-3.23	0.0013	**
ATxGDWFooty	1.77	0.417	4.25	2.12e-05	***
HTAEFooty	0.12	0.106	1.09	0.275	
ATAEFooty	-0.17	0.106	-1.59	0.112	
HTDEFooty	-0.23	0.108	-2.14	0.0322	*
ATDEFooty	0.10	0.107	0.90	0.368	
AvgHTC	0.03	0.035	0.84	0.403	
AvgATC	0.002	0.034	0.06	0.956	
AvgHTCAgainst	-0.02	0.033	-0.73	0.466	
AvgATCAgainst	0.04	0.034	1.07	0.284	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.12: Coefficients Combined FootyStats Model

$$
\begin{aligned}
\log\left(\frac{P(FTR_i = 1)}{1 - P(FTR_i = 1)}\right) = \beta_0 + \beta_1 HomeELO_i + \beta_2 AwayELO_i \\
\beta_3 HTxGASUnder_i + \beta_4 ATxGASUnder_i \\
\beta_5 HTxGDWUnder_i + \beta_6 ATxGDWUnder_i \\
\beta_7 HomexGShotUnder_i + \beta_8 AwayxGShotUnder_i \\
\beta_9 AvgHTCAgainst_i + \beta_{10} AvgATCAgainst_i
\end{aligned}
\tag{4.6}
$$

$$
\begin{aligned}
\log\left(\frac{P(FTR_i = 1)}{1 - P(FTR_i = 1)}\right) = \beta_0 + \beta_1 HomeELO_i + \beta_2 AwayELO_i \\
\beta_3 HTxGASFooty_i + \beta_4 ATxGASFooty_i \\
\beta_5 HTxGDWFooty_i + \beta_6 ATxGDWFooty_i \\
\beta_7 HTDEFooty_i + \beta_8 ATDEFooty_i
\end{aligned}
\tag{4.7}
$$

Chapter 5

Betting

Football and sports betting in general have been rising in popularity for the last several years (H2 Gambling Capital (2022)). Especially, in Europe where football is the most popular sport by far, and in the United Kingdom dominates the whole betting market as 1 billion pounds is bet on football every year in the UK (Gambling Commission UK (2022)). Thus, the betting market is heavily focused on football matches as there is the highest potential profit for the companies. Firstly, this chapter explains the betting logic to introduce the reader to the world of betting, as odds and betting margins need to be understood as the next part of this chapter explains which betting strategies will be applied to what set of matches.

5.1 Betting Market

Even though betting is based on probabilities, the betting sites offer odds, which are easier to understand for bettors and they also directly show how much money the bettor would win if he wagered money on a particular match with particular odds.

Odds	Bet Amount	Implied Probability	Potential Return	Profit
1.5	10	66.7%	15	5
2	10	50%	20	10
3	10	33.3%	30	20

Table 5.1: Betting Odds Explanation

In Table 5.1, where there are three bets placed with the same amount of money on each bet, we can illustrate how this works. On the betting sites, the bettor can see the odds. We can see that in the third column is the implied probability from odds, which is just the inverse of odds. So if the bettor believes the probability of an outcome is 50% the corresponding odds should be 2.00. This leads to a return from betting, which corresponds to:

$$PotentialReturn = BetAmount \times Odds,$$

where *PotentialReturn* corresponds to the money the bettor will receive if his bet was correct including the bet amount he wagered. *BetAmount* is the amount of money the bettor has placed on a bet and *Odds* are the odds offered by the bookmaker on an outcome he decided to bet. Moreover, two important values can be calculated and most of the time are the only two values that the bettor is interested in. Firstly, we calculate profit from a single bet:

$$Profit = BetReturn - BetAmount,$$

where we simply subtract the amount that was wagered from the return the bet generated. If the bet is lost, the potential return is 0 and the profit is suddenly a loss. Most importantly to indicate how efficient is the bettor's betting strategy over a certain period, we calculate the Return on Investment (ROI). In betting, the ROI is used to help the bettor understand if their betting approach is yielding a profit over time. The formula for ROI in betting remains similar to its general use:

$$ROI = \left(\frac{\text{Total Profit/Loss}}{\text{Total Amount Bet}}\right) \times 100,$$

where the Total Profit/Loss is the sum of profit/loss from all the best placed in the period the bettor is interested in and the Total Amount Bet is the amount of money that he bet during this period. So for example, if the bettor placed 100 bets and the total profit was 100$ and he wagered 1500$ in total, the ROI would be:

$$ROI = \left(\frac{\text{Total Profit/Loss}}{\text{Total Amount Bet}}\right) \times 100 = \frac{100}{1500} \times 100 \approx 6.67\%$$

Even though it is important for betting sites to attract as many people to bet on their sites, they also need to ensure that they will be profitable. Thus, the betting margin sometimes referred to as overround is used by betting sites. It can be shown in an example of tossing a coin. Imagine you would bet on the coin toss, which is an event with exact probability of 50%. However, before the toss, the bookmaker would offer you odds such as 1.95 for both outcomes. If we calculate implied probabilities from these odds:

$$\text{Probability Heads} = 1/Odds_H = 1/1.95 \approx 0.513 = 51.3\%$$
$$\text{Probability Tails} = 1/Odds_T = 1/1.95 \approx 0.513 = 51.3\%,$$

we can also sum the probabilities to get approximately 102.6% in total, which means that the extra 2.6% represents the bookmaker's margin. This implies that for every 100$ wagered by the bettor, the bookmaker is expected to pay out around 97.4$, keeping the rest for the betting site. This means that there is no opportunity for profitable wagering, which is confirmed in a study from (Sauer (1998)). However, as in any market, the variety of betting sites offers slightly differing opinions on the likelihood of each outcome, suggesting a divergence in odds among bookmakers confirmed by Constantinou & Fenton (2013), who found that comparing odds across multiple betting sites can be an effective approach to mitigate the bookmaker's margin. This strategy is nowadays even more common as the emergence of software makes it easier to spot these opportunities. Two of the most popular websites for checking maximal possible odds are OddsChecker and OddsPortal[1], where you can compare odds from over 80 leading bookmakers. As mentioned in Section 3.1, our datasets gathered from Football-Data.co.uk[2] contain these maximal odds from season 2019/2020. The datasets contain two types of maximal odds, the total maximal odds, which are the maximal odds that were possible to bet on from the listing of the odds to the start of the match. For example, if the match is played on April 10th and the first bookmaker listed the odds for the match on April 1st, the maximal odds are the ones with the highest value that were available any time between April 1st and April 10th. The maximal closing odds are the ones that are available a few minutes before the start of the match. For different betting sites, the closing odds are available usually 5-10 minutes before the start of the match up until the start of the match, when the match moves to section live betting. Even though the maximal closing odds are slightly lower than the overall maximal odds, they will be used in this study as the use of such odds can be replicated in reality as we know when these odds will be available.

[1]www.oddsportal.com, www.oddschecker.com
[2]https://www.football-data.co.uk/data.php

These odds are in our dataset called MaxCH (maximal closing odds for a home team win) and MaxCDA (maximal closing odds for a draw/away team win). We will test our models on both maximal closing odds and odds from bet365[3], which is one of the biggest sites in the world for sports betting.

5.2 Betting Strategies

The usefulness of the predictive model can be easily diminished by not using the correct betting strategy that suits the model. As there are no studies where betting strategies are discussed while using xG statistics as predictors, we will have to use several betting strategies used typically in football to properly test which one is the most fitting for our models. As we are always predicting the upcoming game week, which was explained in Section 4.3, we will also bet on each game week separately to allow us to account for the data in the current season to achieve the best results possible. We have a binary outcome variable (either a home team winning or an away team not losing), we will bet either a home team win or an away team not losing. In the betting world, betting on the away team not losing the match is called Double Chance, which is the possibility to bet on two out of possible three outcomes. From the regression, we obtain a vector of predicted probabilities $predictions_i$ that the binary outcome variable (in our case FTR) is 1. These probabilities are then converted to odds as follows:

$$OddsHomeWin_i = \frac{1}{predictions_i},$$
$$OddsDrawAwayWin_i = \frac{1}{1 - predictions_i},$$

where $OddsHomeWin_i$ and $OddsDrawAwayWin_i$ are the odds for each match in the test set indexed by the index i. Betting will be divided into two sections based on the strategy of allocating our betting budget. Firstly, we will use a simple strategy, which is betting on every match based on the odds gathered from the models:

$$\text{Bet}_i = \begin{cases} \text{Home Win}, & \text{if } OddsHomeWin_i < OddsDrawAwayWin_i \\ \text{Draw/Away Win}, & \text{if } OddsHomeWin_i > OddsDrawAwayWin_i \end{cases}$$

The second strategy will consist of finding "value bets", which arises when the bettor believes that the probability of the outcome is higher than the book-

[3]www.bet365.com

maker's. Implying that the odds offered by bookmakers are higher than the expected odds by bettors. The value bet can be illustrated by the expected value.

$$EV_{Bet} = (P_{\text{win}} \times \text{Total Profit}) - (P_{\text{lose}} \times \text{Bet Amount})$$

For example, if the odds offered by the bookmaker are 2.00 and you bet 1\$. If you win the bet, you receive 2\$ in total (including the 1\$ you placed on a bet). If you lose, you simply lose the 1\$. Odds 2.00 mean that the estimated probability of an outcome by a bookmaker is 50%. If the bettor believes that the probability of the outcome is for this example 55%, the expected value of this bet is as follows:

$$EV_{Value} = (0.55 \times 1\$) - (0.45 \times 1\$) = 0.1\$$$

Thus, the expected value of betting on a match, where the bettor believes the probability is higher by 5%, is 0.1\$. This means that in the long run, the bettor is expected to make an average profit of 0.1\$ if his estimated probability is accurate. For this strategy, we will compare the $predictions_i$ that our models have estimated with the predictions of the bookmakers to find the matches with the possibility of a value bet. We will be referring to these matches as mispriced matches. The process of recognizing mispriced matches by our models can be summarized in the following steps:

- First we have to calculate the probability of the away team not losing the match.
- We have to calculate the probabilities implied from the bookmaker's odds MaxCH and MaxCDA.
- Set threshold θ for indicating mispriced matches.

As our outcome (FTR) is a binary variable, and the probabilities stored in the vector $predictions_i$ are probabilities that $FTR = 1$ (the home team wins), we calculate the probability of an away team not losing the match using the symmetry of logistic regression:

$$ProbH_i = predictions_i,$$
$$ProbDA_i = 1 - predictions_i,$$

where $ProbH_i$ is the probability of a home team winning the $match_i$ and $ProbDA_i$, is the probability of the away team not losing the $match_i$ (draw/away win). As odds are only inverse of probabilities, we get the implied probabilities from the bookmaker's odds as follows:

$$MaxProbH_i = 1/MaxCH,$$
$$MaxProbDA_i = 1/MaxCDA,$$

where $MaxProbH_i$ is the implied probability from maximal closing odds of a home team winning the $match_i$ and $MaxProbDA_i$, is the implied probability from maximal closing odds of an away team not losing the $match_i$. Thus, as explained above the $match_i$ is identified as mispriced if either of the following conditions is met:

$$\begin{cases} ProbH_i - MaxProbH_i > \theta \\ ProbDA_i - MaxProbDA_i > \theta \end{cases}$$

Thus, we will have two sets of matches to which we will apply our betting strategies explained later. Moreover, we will bet on two seasons, particularly seasons 4 and 5 in our dataset, which is the seasons 2021/2022 and 2022/2023, as the results are slightly different each season and the ratio of home team wins and away team not losing is not the same every season according to Table 3.8. We will be most interested in calculating the overall ROI as it evaluates the efficiency and profitability of an investment (in our case betting strategy).

5.2.1 Fixed Betting

For our study, we will use several money management techniques that were mentioned by van Wijk (2021) and Hubáček *et al.* (2019). The first technique for deciding how much to bet on each match is called fixed betting strategy, which is a simple and most used technique in the betting world. A fixed amount is placed on each bet without any influence of other factors. For example:

- Let the bettor have 1000$ as his budget for betting.
- The only criterion he has to consider is not to pick a value high enough to make him lose the whole budget in a few unfortunate bets.
- Imagine the bettor decides to bet 10$ on each bet.

Then in this example he has 100 bets that he can wager without worrying about losing a whole budget and if his ability to predict outcomes of matches is more accurate than the predictions of bookmakers, he should make a profit in the long run. We can illustrate how the strategy works on an example in Table 5.2, where 5 bets have been placed and we can see the odds for each bet and the total profit/loss.

Outcome	BetAmount ($)	Odds	Total Profit/Loss ($)
Win	10	2.0	+10
Loss	10	3.2	-10
Win	10	2.5	+15
Loss	10	1.8	-10
Win	10	3.0	+20

Table 5.2: Fixed Betting Overview

This approach simplifies the management of the betting budget and mitigates the risk of significant losses as the amount to lose is always 10$. However, it is limited to potential returns as neither the odds nor estimated likelihood by the bettor is a factor that influences the bet amount. The profit/loss from a single bet can be calculated as follows:

$$\text{Profit/Loss} = \begin{cases} BetAmount \times (Odds_{BET}), & \text{if the bet was correct,} \\ -BetAmount, & \text{if the bet loses,} \end{cases}$$

where $Odds_{BET}$ are the odds for the particular bet. Similarly, the return on investment can be calculated after n number of bets:

$$\text{ROI} = \frac{\sum_{i=1}^{n} \text{Profit/Loss}_i}{n \times BetAmount} \times 100,$$

where the summarized $Profit/Loss$ from all bets from i to n is divided by the number of placed bets times the amount of money on a single bet.

5.2.2 Proportional Betting

A similar strategy is proportional betting, where the bettor allocates a percentage of his betting budget on the next bet. This method is simple but allows the bettor to balance growth and risk, unlike the fixed betting strategy, the bet amount increases/decreases based on the results of previous bets.

- Let bettor have 1000$ as his budget on betting at the start
- Let p be the proportion of the budget he decides to put on each bet

We will assume that the bettor decides that he wants to bet 1% of his starting budget. For betting in sports, most matches are played during the weekends on Saturday and Sunday in the evenings, so it would be unrealistic for bettors to recalculate their budget after each bet as many matches are played simultaneously and the outcome of one match is not yet known before the start of next match. Thus, we will assume he places 10 bets on a Friday evening and he will evaluate his results by the end of the weekend.

Amount Bet	Profit/Loss	Current Budget
100	50	1050

Table 5.3: Weekend Bets Result

The total amount that was wagered was 100$ as he placed 10 bets that were 1% of his starting betting budget. For this example, he made a profit of 50$, which means that for the next betting session, his budget will start at 1050$ and the bet amount will go up to 10.5$ a bet.

The calculation for ROI using this strategy is as follows:

$$\text{ROI} = \frac{\sum_{i=1}^{n} Profit/Loss_i}{\sum_{i=1}^{n} BetAmount_i} \times 100,$$

where the sum in the numerator represents the sum of Profit/Loss from each bet during all betting sessions, indexed from i to n divided by the sum of money wagered on all bets. Compared to the fixed betting strategy, the $BetAmount$ cannot be multiplied by the number of bets n as the amount of money wagered on a single bet is different for each betting session.

As our strategy is to bet on each game week separately, we will use the same strategy as illustrated in Table 5.3. Thus, after each game week, the betting budget will be recalculated and we will then bet the proportion of our new budget on the next game week.

5.2.3 Fixed Expected Return Betting

The last simple betting strategy is called the fixed expected return strategy, where the bettor allocates such an amount that the expected return from the bet is always the same amount he has determined at the beginning. The goal of this strategy is to have consistent returns, but it is riskier as betting on more likely outcomes can lead to a larger bet amount. Some bettors also adjust their target win amount to cover for previous losses. For example, if he lost 3 bets in a row, the next time the expected return from all bets aims to cover those losses plus the desired profit.

- Let bettor have 1000$ as his budget on betting at the start.
- Bettor decides that from every bet he wagers, the expected return is 10$.

We can illustrate the strategy on a few bets in the following table:

Odds	BetAmount ($)	Expected Profit ($)
2.0	10	10
3	5	10
1.5	20	10

Table 5.4: Fixed Return Betting Examples to Win $10

We can see that the bet amount always depends on the odds as the targeted profit is 10$. This strategy is a simple way to mitigate risk as for higher odds the bet amount will be lower. On the other hand, this strategy tends to larger stakes when betting on smaller odds so the betting budget needs to be sufficient to cover for these bets. The ROI is calculated as follows:

$$\text{ROI} = \frac{\sum_{i=1}^{n} Profit/Loss_i}{\sum_{i=1}^{n} BetAmount_i} \times 100,$$

where the sum of profit/loss from all bets in the period the bettor is interested in, is divided by the sum of all the money wagered on the bets.

5.2.4 Kelly Criterion

Furthermore, we will also test some advanced techniques such as the Kelly Criterion for betting. The strategy was invented by Kelly Jr. (1956) and is used in both the investment industry but also in sports betting. It is a formula for sizing a bet, found by maximizing the expected value of the logarithm of wealth, which is equivalent to maximizing the expected geometric growth. It is mainly used when the return on investment of in our case outcome of a bet is binary. The Kelly formula (f^*) is defined as follows:

$$f^* = p - \frac{1-p}{b},$$

where p is the probability of a win, $(1-p)$ is the probability of losing sometimes referred to as q divided by b, which is the proportion of the bet gained with a win. For decimal odds, b can be easily calculated:

$$b = Odds - 1,$$

where *Odds* are the odds on which the bettor placed a wager. Thus, we can illustrate how the fraction is calculated in the example with a coin landing on heads with odds of 2.00. However, the coin is biased and has a 52% chance of ending up on heads. Then:

- $p = 0.52$,
- $(1-p) = 0.48$,
- $b = 2.00 - 1 = 1.00$,
- $f^* = 0.52 - \frac{0.48}{1} = 0.04$

Meaning that Kelly Criterion suggests that you should bet 4% of your current budget. Similarly, to previous techniques, we are interested in ROI, which can be calculated as:

$$ROI = \frac{\sum_{i=1}^{n} Profit/Loss_i}{\sum_{i=1}^{n} KellyBetAmount_i} \times 100,$$

where the sum of Profit/Loss from all bets placed is divided by the sum of the amount bet. For the Kelly Criterion strategy, we will simplify the betting strategy. As it calculates the proportion of the betting budget that should be wagered on independent sporting events, we will have a budget of 100 units for each game week and we will bet on each match independently based on the fraction calculated from the Kelly formula.

5.2.5 Fibonacci Sequence

Lastly, we will evaluate the performance of our models using the Fibonacci Sequence, where each subsequent number in this sequence is the sum of the previous two numbers in the sequence (Sigler (2003)). Mathematically, it is defined by the recurrence relation:

$$F(n) = F(n-1) + F(n-2), \qquad \text{for } n > 1$$

In the sense of betting, we define a starting amount of money that will be the first number in the sequence.

- Let Bettor decide to start with 10$

Then if he loses, he moves to the next number in the Fibonacci sequence, so he bets 10$ again. If he loses again, he will bet 20$ as the sum of previous bets is $10 + 10$.

Bet in a row	BetAmount ($)
1	10
2	10
3	20
4	30
5	50
6	80
7	130
8	210
9	340

Table 5.5: Fibonacci Sequence in Betting

However, this table illustrates what would happen if the bettor lost his first 8 bets. If the bettor wins, he moves two numbers back in the sequence for his next bet. If it is his first or second bet, he simply starts over and bets 10$ as his first bet in the sequence.

This approach is popular among bettors as it offers a structured method for managing bets, which means that after a series of losses, the bettor can systematically mitigate his loss instead of making rash decisions. However, it has several drawbacks. First of all, it does not require many lost bets in a row to reach large bet sizes as we can see in Table 5.5, where a bettor that starts with a 10$ bet is supposed to bet 340$ if the first 8 bets were losing. There are two limitations connected to large stakes, the first is the betting budget, which for most bettors is strictly limited. Additionally, even if the bettor had an unlimited amount of money, the vast majority of betting sites have maximum bet limits, which can prevent the Fibonacci sequence from working as the bettor would eventually have to deal with the site denying his bet.

As for previous strategies, the ROI is calculated analogously:

$$\text{ROI} = \frac{\sum_{i=1}^{n} Profit/Loss_i}{\sum_{i=1}^{n} BetAmount_i} \times 100,$$

by dividing the sum of profit/loss from all bets by the total bet amount. The Fibonacci sequence is a complicated approach for us, as we do not always know the result of our bet before the time we want to bet on the next match result. Thus, we will assume that we do know it to test the strategy. The sequence of bets will constructed analogously to Table 5.5, so the starting bet will be 10 units.

Chapter 6

Results

6.1 Accuracy

Firstly we will look at the prediction accuracy of our models. Accuracy is the most important evaluation metric when it comes to understanding the performance of a model using logistic regression. It is defined as the proportion of true results (both true positives and true negatives) among the total number of cases examined. In our case, this means that it is the proportion of correctly predicted outcomes of matches divided by the total number of matches that we predicted. The general formula for calculating accuracy is following:

$$\text{Accuracy} = \frac{\text{Total Number of Correct Predictions}}{\text{Total Number of Predictions}},$$

Thus,

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

where:

- TP = True Positives: The model predicts that a home team wins as the outcome ($FTR = 1$) and the outcome of a match is indeed a home team win ($FTR = 1$).

- TN = True Negatives: The model predicts that an away does not lose as the outcome ($FTR = 0$) and the outcome of a match is indeed an away team not losing ($FTR = 0$).

- FP = False Positives: The model predicts that a home team wins as the outcome ($FTR = 1$), but the actual result of the match was an away team not losing ($FTR = 0$).

- FN = False Negatives: The model predicts that an away does not lose as the outcome ($FTR = 0$), but the actual result of the match was a home team win ($FTR = 1$)

By definition, the accuracy ranges between 0 and 1. A value of 0.5 suggests that the model does no better than taking a guess, whereas values closer to 1 indicate higher effectiveness than just a coin flip. Accuracy is widely used because it is easy to calculate, easy to interpret, and can be used for datasets of any size. However, if the ratio of outcome variables is misbalanced in the dataset, the accuracy metric yields misleadingly high performances that result from systematically predicting the majority class. In our dataset, the ratio of more occurring outcomes, which is an away team not losing ($FTR = 0$), is around 56%, which means that we aim to surpass this and achieve accuracy higher than 56%.

Model	Season 4 (%)	Season 5 (%)	Overall Accuracy (%)
Traditional	66.74	63.82	65.28
xG Under	66.35	63.17	64.76
xG Footy	65.23	63.35	64.29
Combined Under	65.97	64.44	65.21
Combined Footy	66.62	64.14	65.38
Combined Under S	66.58	64.64	65.61
Combined Footy S	66.83	63.86	65.35
Overall Accuracy	66.33	63.92	65.13

Table 6.1: Model Accuracy Across Seasons 4 and 5

The Table 6.1 represents the accuracies of all models across the two seasons, which we decided to test our models on (Combined Under S and Combined Footy S are the combined models including only significant variables from the original combined models). Firstly, we can see that including only significant variables improved the accuracy of the Combined Understat model across both seasons. The overall accuracy was higher by 0.4%. However, for the Combined FootyStats model, we can see that the accuracy dropped by 0.03%. Importantly, all models overcame the desired threshold of 56% for both seasons. We can see that the accuracy of all models combined for Season 4 was higher than the accuracy for Season 5 by 2.41%. The highest accuracy in one season was achieved by the Combined FootyStats model, where the accuracy was approximately 66.83% in season 4. The poorest performance was from the model based on solely the xG variables calculated from Understat xG in Season 5 (63.17%). We can see that the overall accuracies for both seasons combined were above 65% for the Traditional Model and the combined models. The highest overall accuracy had the Combined Understat model including only significant variables, reaching an accuracy of 65.61%. The accuracy of all models across both seasons was approximately 65.13%.

6.2 Betting Results

The main goal of this thesis was to test the performance of the xG metric while betting. As explained in Chapter 5, we tried two different approaches using several money management techniques, first betting on every match, and second betting only on mispriced matches. This section provides tables and graphs demonstrating how these strategies worked for each model and how well the models including xG variables performed compared to the Traditional Model.

6.2.1 Betting on Every Match

Firstly, we look at the results of betting on every match in Seasons 4 and 5. This strategy was simple, we always bet on the team that was more likely to win based on the probabilities according to our models. As explained in Section 5.1, firstly the maximal closing odds were used to achieve the highest profit possible. We are interested in a return on investment, illustrated in the following tables.

Model (Values in %)	FIX	PROP	FER	KEL	FIB
Traditional	-0.04	-0.25	-1.28	5.29	0.39
xGUnder	1.84	1.52	-0.64	4.88	1.48
xGFooty	-0.32	-0.55	-1.35	3.57	-0.75
CombinedUnder	-1.24	-1.39	-1.64	1.83	-0.83
CombinedUnderS	-0.18	-0.40	-1.30	2.50	-0.06
CombinedFooty	-0.20	-0.43	-1.31	4.34	-0.01
CombinedFootyS	0.07	-0.18	-1.23	3.37	0.13

Table 6.2: Model's ROI Season 4 (Maximal Closing Odds)

Model (Values in %)	FIX	PROP	FER	KEL	FIB
Traditional	-4.76	-4.09	-6.10	-4.14	-6.60
xGUnder	-4.28	-4.30	-6.05	3.47	-3.04
xGFooty	-4.43	-4.38	-6.10	-1.15	-6.24
CombinedUnder	-3.56	-3.24	-5.73	-0.99	-2.36
CombinedUnderS	-3.06	-2.60	-5.58	-0.63	-2.83
CombinedFooty	-4.29	-3.63	-5.97	-4.46	-3.05
CombinedFootyS	-4.72	-3.95	-6.12	-5.07	-3.44

Table 6.3: Model's ROI Season 5 (Maximal Closing Odds)

Firstly, Tables 6.2 and 6.3 present the returns on investment in each season separately. We can see that the betting results correspond to the accuracies represented in Table 6.1 as the results are mostly negative for Season 5. If we look at the results for Season 4, it is clear that the Kelly Criterion (represented by KEL in the table) was the most efficient strategy as all models are showing positive results and the average ROI of the strategy across all models is approximately 3.68%. On the other hand, the Fixed Expected Return strategy (FER) was the worst as none of the models had positive results using this strategy and averaged -1.25% ROI across all 7 models. When looking at the first two strategies, which are Fixed Betting (FIX) and Proportional Betting (PROP), we examine slightly better results for Fixed Betting. Lastly, the Fibonacci sequence (FIB) seems to be the second-best strategy with the Traditional, xGUnder, and CombinedFootyS showing positive values of ROI. The best-performing model across all strategies was the model using only xG variables created from Understat xG values, with an average ROI of around 1.82%. Table 6.3 representing ROI for Season 5, we again conclude that the Kelly Criterion was performing the best compared to the other strategies. However, the average ROI for all 7 models was around -1.85% which is still a poor performance. The worst strategy was the Fixed Expected Return, as all models exhibited an ROI of -5.5% and more. The only positive ROI in Season 5 was by the best-performing model in Season 4 also, which is the Understat xG model using the Kelly Criterion strategy. This means that the model based only on variables from xG statistics from Understat managed to be profitable in both seasons using this strategy.

This is confirmed in Table 6.4 which shows the average ROI from Season 4 and Season 5 together. The Understat xG model had 4.18% ROI across both seasons while using the Kelly strategy, which is the best result by 2.97%

Model (Values in %)	FIX	PROP	FER	KEL	FIB
Traditional	-2.40	-2.17	-3.69	0.58	-3.11
xGUnder	-1.22	-1.39	-3.36	4.18	-0.78
xGFooty	-2.38	-2.47	-3.73	1.21	-3.49
CombinedUnder	-2.40	-2.32	-3.68	0.42	-1.60
CombinedUnderS	-1.62	-1.50	-3.44	0.94	-1.45
CombinedFooty	-2.25	-2.03	-3.64	-0.06	-1.53
CombinedFootyS	-2.33	-2.07	-3.68	-0.85	-1.66

Table 6.4: Model's ROI in Both Seasons Together (Maximal Closing Odds)

compared to the ROI of the FootyStats xG model that achieved 1.21% ROI using the Kelly Criterion. All models except for the two combined FootyStats models managed to achieve positive results with this strategy. Importantly, the Understat xG model was able to have better results than the Traditional model across both seasons using all 5 strategies and the FootyStats model was better with the Fixed Betting strategy and mainly the Kelly Criterion strategy, where the results of betting were more profitable by 0.63%. Additionally, the combined models were able to outperform the model using solely the traditional variables using all strategies except for the Kelly Criterion, where both models combined from the traditional variables and FootyStats xG variables, had worse results.

Moreover, our dataset also contains the betting odds from a single bookmaker. Thus, the following Table 6.5 contains the ROI when betting on odds offered by bet365[1] using the Kelly Criterion strategy as it was the only strategy that was consistently profitable using maximal closing odds. It would not make sense to use betting strategies that were not profitable for even higher odds. The table is again divided into the Season 4 and Season 5 given that the two seasons generate slightly different results. Similarly, to the maximal closing odds we observe quite a significant difference in the ROI of each season. Firstly, we see that the Understat xG model achieved positive numbers in both seasons, even though the profit was very small. All models except the combined Understat models were profitable in Season 4, which corresponds to the fact that the profit in Season 4 while using the maximal closing odds was the smallest for these two models as well (when using the Kelly Criterion). The highest ROI was 2.35% in Season 4 for the Combined FootyStats model, but it is compensated

[1]www.bet365.com

Model (Values in %)	Season 4 ROI	Season 5 ROI	Average ROI
Traditional	1.30	-5.18	-1.94
xGUnder	0.49	0.21	0.35
xGFooty	0.57	-4.99	-2.21
CombinedUnder	-2.20	-2.82	-2.51
CombinedUnderS	-1.46	-2.71	-2.09
CombinedFooty	2.35	-6.64	-2.15
CombinedFootyS	0.60	-7.54	-3.47

Table 6.5: Model's ROI Using Kelly Criterion (B365 Odds)

by the poor performance in Season 5. Interestingly, the Traditional Model had better overall results than all combined models, which was not the case when using the maximal closing odds. To sum it up, the only profitable model using the B365 odds was the Understat xG model, otherwise, the ROI ranges between -3.47% to -1.94%.

6.2.2 Betting on Mispriced Matches

The second strategy was to find mispriced matches based on the implied probabilities from the bookmaker's odds and the probabilities of our models as explained in Section 5.2. The threshold for the match to be mispriced was set to 0.1, which means that the match is mispriced if our model believes that either of the outcomes is more likely to happen by 10% and more than the implied probability from the bookmaker's odds. As in the previous section, we want to maximize our profit and achieve the highest return on investment possible, so we use the maximal closing odds first.

First of all, it is important to mention that the Fibonacci sequence was not used at the end for betting on mispriced matches as the bets were growing to unrealistic numbers for betting in reality. The problem was that the range of matches that each model indicated as mispriced was between 201 and 480 matches with average odds from 2.97 to 3.74. This means that the models were more inclined towards outcomes with smaller probabilities, which led to the sequences of many lost bets in a row, which is a drawback of using the Fibonacci sequence.

Table 6.6 shows the results for all 4 strategies (without the Fibonacci sequence) in Season 4. We can see that the results improved significantly as all strategies are profitable apart from the combination of the Proportional Betting strategy and the Traditional model, where the ROI was -0.18%. The best results are again for the Kelly Criterion strategy as the average ROI for all models was approximately 8.07% and the highest ROI from all strategies when betting based on the Combined FootyStats model, which yielded a 13.1% return on investment from betting on 234 matches from the whole season with average odds of 3.41. For illustration, when converted to probabilities, the average likelihood of outcome that we bet on in this combination of model and strategy was around 29.3%. Out of all the bets, 91 of them were correct and 143 were wrong, which means that the success ratio was only 38.9%. However, this is balanced by the higher odds that we were betting on and also by the Kelly Criterion which decides what portion of the betting budget to allocate for each bet.

Model (Values in %)	**FIX**	**PROP**	**FER**	**KEL**
Traditional	0.74	-0.18	8.28	6.99
xGUnder	4.48	4.11	4.32	5.50
xGFooty	4.97	4.19	5.00	4.46
CombinedUnder	6.91	6.12	7.71	8.46
CombinedUnderS	4.85	4.2	6.95	6.80
CombinedFooty	10.2	9.21	10.3	13.1
CombinedFootyS	7.14	6.23	7.40	11.4

Table 6.6: Models ROI Season 4 - Mispriced Matches (Maximal Closing Odds)

The starting budget was 100 units for each game week as explained in Section 5.2 and the average bet size for one bet was approximately 24.16. The amount bet during the whole season was 5653.81 units and the net profit was 740.33. The highest odds in the whole season were 17.00 and the highest odds for a successful bet were 8.64. However, the smallest odds were 1.32, so the model did not indicate only matches with higher odds as mispriced. Moreover, we can see that the combined models are again outperforming the Traditional model using all betting strategies with only one exception and that is the Fixed Expected Return strategy, where neither of the two combined Understat models is outperforming the Traditional model. However, for the Fixed Betting strategy and Proportional Betting strategy, we can see that the difference is more than 6%. On the other hand, we can see that the models based purely on xG variables slightly struggled along with the Traditional model compared to the combined ones. The average ROI was around 4.62% for both models across all strategies, which is still a significant number considering that the Understat xG model indicated 468 matches as mispriced in the whole season, and 480 matches were pointed out by the FootyStats xG model. We observe that the Fixed Betting, Proportional Betting, and Fixed Expected Return Betting strategies have very similar results except for the Traditional model which has much better results with the Fixed Expected Return approach. This might be caused by the higher accuracy of predicting the results of more likely outcomes. Out of the 95 matches that were predicted correctly, 68 of them had lower odds than 3.00. Additionally, 50 lost bets were with odds above 4.00, where the losses are controlled by the amount bet, which is $\frac{1}{4}$ of bet size in Fixed Betting for odds 4.00. Finally, the least performing model was the Traditional model with an ROI of around 3.96% across all strategies. If looking at the results from Season 5 represented by Table 6.7, we can see that the results are again worse, but we are still mainly in positive values. The Kelly Criterion strategy can be concluded as the best betting strategy for all of our models as also for the mispriced matches it has the best results by far for both seasons.

Model (Values in %)	**FIX**	**PROP**	**FER**	**KEL**
Traditional	5.55	4.66	-1.87	5.78
xGUnder	-0.25	-1.96	4.35	4.27
xGFooty	-3.58	-5.55	0.41	0.13
CombinedUnder	2.14	1.49	5.17	8.29
CombinedUnderS	5.30	4.37	6.78	9.14
CombinedFooty	3.87	3.55	5.16	8.63
CombinedFootyS	0.15	-0.44	5.52	5.98

Table 6.7: Models ROI Season 5 - Mispriced Matches (Maximal Closing Odds)

Firstly, we examine that both models constructed only from xG variables are in positive values for at least two strategies in both seasons unlike when betting on every match, where the FootyStats xG model had -1.15% ROI in Season 5 using the Kelly Criterion strategy. Even though the Traditional model has more balanced results with combined models this season, it is still outperformed with Fixed Expected Return strategy and Kelly Criterion strategy where the difference between the best results of the Traditional model and the two combined models is around 3%. The best results exhibit the Combined FootyStats model along with the Combined Understat model with only significant variables using the Kelly Criterion, where both have above 8.5% and the Understat model has

9.14% ROI. We will also look at Table 6.8 representing the ROI across both seasons for all models. Firstly, it is important to acknowledge that both xG models were profitable across both seasons apart from the FootyStats model while using the Proportional Betting strategy. The highest ROI was 4.89% for the xG Understat model using the Kelly Criterion strategy, but for the FootyStats model, the best result was surprisingly while using the Fixed Expected Return strategy that yielded 2.71% ROI. The total number of matches that were indicated as mispriced was 821 for the Understat model and 831 for the FootyStats model.

Model (Values in %)	FIX	PROP	FER	KEL
Traditional	3.15	2.24	3.21	6.39
xGUnder	2.12	1.08	4.34	4.89
xGFooty	0.70	-0.68	2.71	2.30
CombinedUnder	4.53	3.81	6.44	8.38
CombinedUnderS	5.08	4.29	6.87	7.97
CombinedFooty	7.04	6.38	7.73	10.87
CombinedFootyS	3.65	2.90	6.46	8.69

Table 6.8: Model's ROI in Both Seasons Together - Mispriced Matches (Maximal Closing Odds)

For the mispriced matches, the best result overall achieved the Combined FootyStats model with 10.87% ROI when using the Kelly Criterion betting strategy. When compared to the Traditional model, it is clear that the overall performance was significantly better. The Traditional model achieved an average ROI of 3.75% across all 4 betting strategies while betting on 469 matches in total and the Combined FootyStats model achieved 8.01% ROI using all strategies and the number of total matches wagered was 463. Moreover, the Combined Understat model (significant variables only) was also performing well compared to the Traditional model, with 6.05% profit during both seasons with all betting strategies. Lastly, in Table 6.9, illustrating the ROI using the Kelly Criterion strategy for mispriced matches, when betting only on odds offered by B365 as in Section 6.2.1. First of all the threshold has been lowered to 0.8 (8% difference

Model (Values in %)	Season 4 ROI	Season 5 ROI	Average ROI
Traditional	-1.17	3.84	1.34
xGUnder	-6.53	-1.87	-4.20
xGFooty	-2.60	-6.98	-4.79
CombinedUnder	-4.40	6.74	1.17
CombinedUnderS	-3.45	4.16	0.35
CombinedFooty	0.56	2.37	1.46
CombinedFootyS	1.89	-1.11	0.39

Table 6.9: Model's ROI Kelly Strategy - Mispriced Matches (B365 Odds)

in estimated probabilities) As with the original threshold, some models indicated less than 100 matches as mispriced. By lowering the threshold, the lowest number of matches was 170 for the Combined FootyStats model in Season 5. From the table, we can see that it was also the only model that achieved positive ROI in both seasons. It also had the highest average ROI of 1.46% from a total of 353 games across both seasons, followed by the Traditional model with 1.34%. The Combined Understat model was the third most profitable model when betting on B365 odds with 1.17% profit, but it was from the most matches as the model marked 414 matches as mispriced. The xG models were not able to be profitable as both indicated nearly 800 matches as mispriced, which is approximately twice the amount of other models.

To sum it up, the most certain finding is that the Kelly Criterion is by far the best betting strategy for all models. Additionally, the xG models performed exceptionally well when betting on every match using the maximal closing odds and only decided the outcome to bet. The model based on variables calculated from Understat xG values was the best-performing by 3% in ROI, followed by the model based solely on xG variables from FootyStats. Both managed to be profitable across both seasons using the Kelly Criterion and the difference between the ROI of the Traditional model and the Understat xG model was over 3.6%. The Understat xG model was profitable in both seasons when betting on every match using only the B365 odds with 0.49% ROI in Season 4 and 0.21% in Season 5. Even though, the profit is nearly 0, beating the bookmaker is a significant sign of the predictive power of xG variables. Moreover, we show that the combination of traditional statistics and xG statistics can improve the predictive power of the model using only traditional statistics. When betting only on mispriced matches, the combined models appear to be even stronger as they were the most profitable ones across both seasons followed by the Traditional model. Combined with Kelly Criterion, the xG models improved only slightly with ROI rising by approximately 1% for both models. However, unlike betting on every match, the xG models were profitable using all strategies except the Proportional Betting strategy, where the FootyStats xG model had a negative return on investment. Similarly to the betting on every match, the best result of the Traditional model was 6.39% ROI with the Kelly Criterion strategy, which was still 4.48% worse than the Combined FootyStats model which achieved 10.87% ROI.

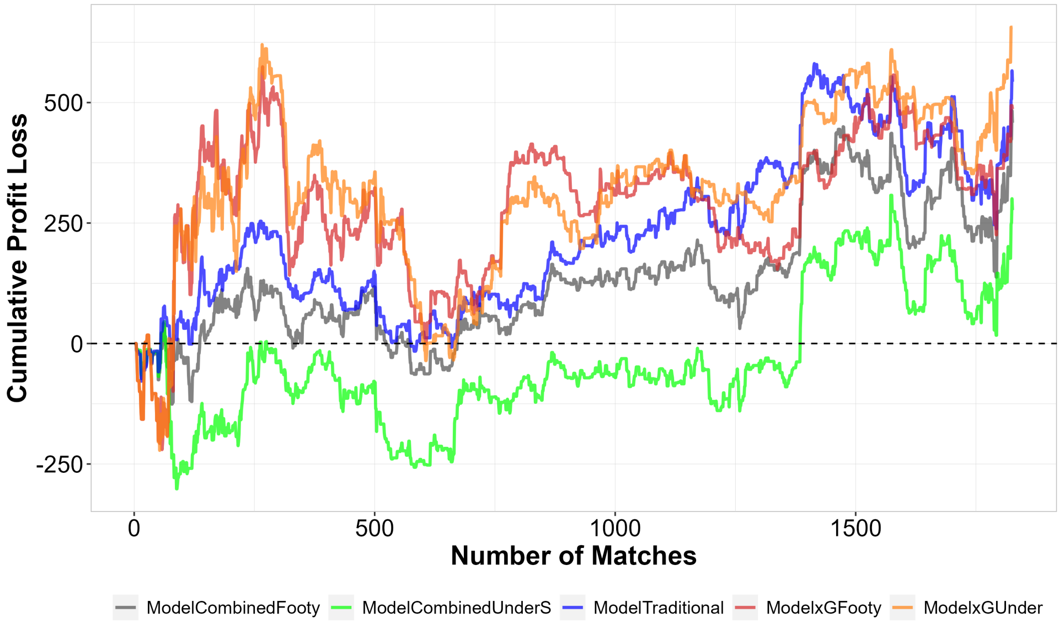

Figure 6.1: Kelly Criterion Season 4 (Maximal Closing Odds)

Furthermore, we can look at the profit/loss, which does not consider the bet amount during the whole period. Figure 6.1 represents the cumulative profit/loss development in season 4 using the Kelly Criterion strategy for betting. We observe Traditional, both xG models, and the two combined models with higher ROI from Table 6.2. The graph demonstrates that at the beginning of the season, all models were slightly at a loss, but after approximately 100 matches all models except for the Combined Understat model became profitable and remained during the whole season. The Combined Understat model managed to become profitable during the 29th game week (of 38 total game weeks). Whereas, before the start of the game week, the model had a loss of approximately -50.49 units, after the game week it had a profit of 169.42 units. The graph also suggests that the xG models were performing well at the start of the season, where they are well above the rest of the models. The highest cumulative profit was achieved by Understat xG model (656.38 units) at the end of the season while having a lower ROI than the Traditional. This is caused by the fact, that the total bet size for all 1826 matches in the season was 13454.11 units for the xG model and only 10267.31 for Traditional. Nevertheless, season 5 had more interesting results as the Understat xG model was the only model that remained in the positive numbers. This is confirmed in Figure 6.2 representing the development of profit/loss in season 5, where the graph depicts the strong performance of the xG model. Unlike the previous season, where the Combined Understat model was the worst performing during the whole season, this season it was competing with the xG model and remained profitable or around zero until the 36th game week.

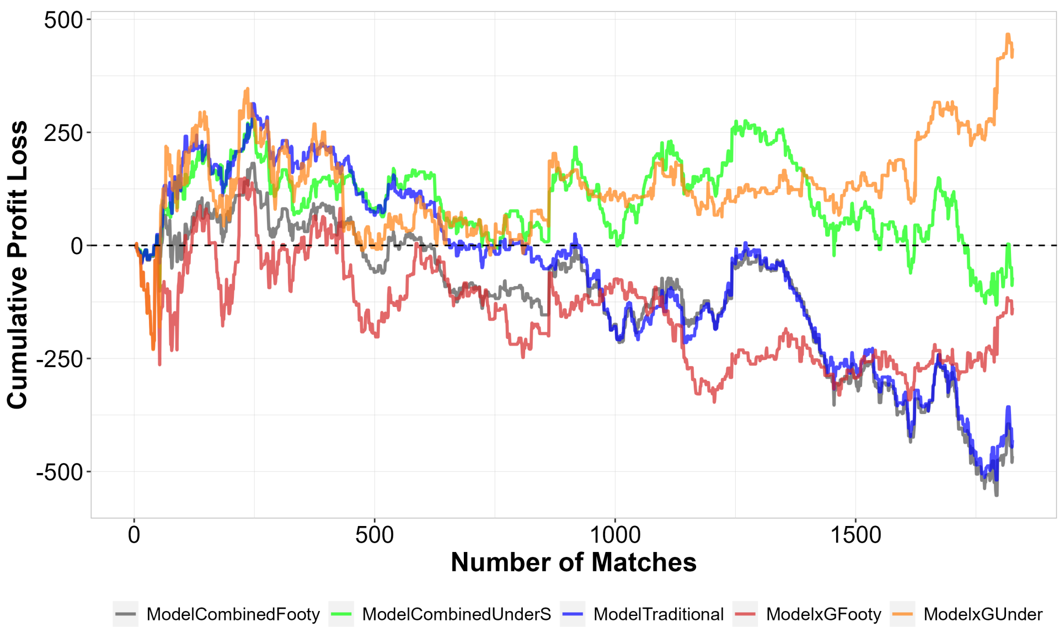

Figure 6.2: Kelly Criterion Season 5 (Maximal Closing Odds)

The same pattern can be seen in Figure 6.3, which demonstrates the profit/loss using the bet365 odds in season 5. Similarly to Figure 6.2, the Combined Understat model is at equal or better numbers for nearly the whole season compared to the xG model. We can see that the Understat xG model performed well at the beginning of the season, having its peak profit after the 5th game week (278.79 units). On the other hand, the FootyStats xG model was performing poorly and it was in non-profitable values since game week 8 and continued to lose more until reaching the peak loss of -604.71 units by the end of the 33rd game week. The Understat xG model ended the season with a cumulative

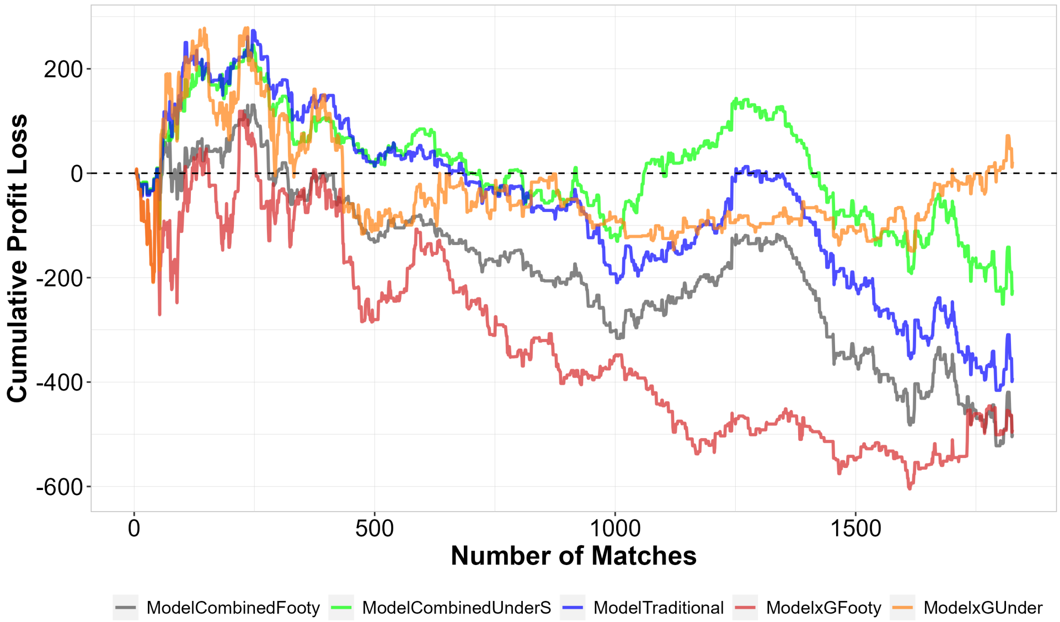

Figure 6.3: Kelly Criterion Season 5 (B365 Odds)

profit of 71.95 units after betting a total of 10300.53 units, while the Traditional model had a cumulative loss of -392.27 units with a total bet size of 7571.26 units. The illustration of cumulative profit/loss for mispriced matches varies because each model has a different number of matches that it identified as mispriced. Consequently, the cumulative profit curves have different lengths, which is a characteristic of the data and not a misrepresentation by the graphical visualization. Looking at Figure 6.4 illustrating the Kelly Criterion betting strategy in season 5, we can see that as mentioned above, the Traditional and combined models indicate only slightly above 200 matches as mispriced, while the xG models indicate around 350 games. This means that even though by the end of the season the cumulative profit of the Understat xG model (415.65 units) was higher than the profit of the Traditional model (309.89), the ROI was higher for the Traditional model (Table 6.7). This is caused by the bet amount, which was again higher for xG models as they suggested betting on a greater portion of matches. However, we can draw other conclusions from the graph such as that except for the first few matches both Combined Understat and Traditional models were profitable during the whole season, while the Combined FootyStats profit hovered around zero for approximately 60 matches and only after that got to more promising numbers, eventually overcoming the profit of Traditional model. On the other

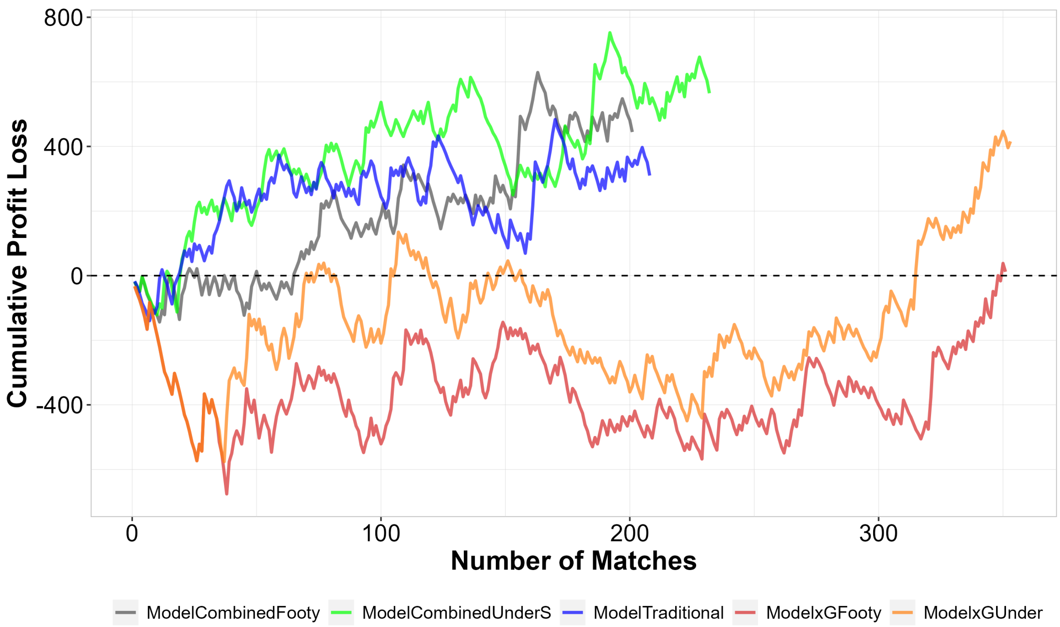

Figure 6.4: Kelly Criterion Season 5 - Mispriced Matches (Maximal Closing Odds)

hand, the xG models were mostly in negative values and managed to climb to positive numbers by the end of the season, which is a similar trend that was visible in Figure 6.2 and Figure 6.3. In conclusion, the profit/loss graphs can be a useful tool for indicating trends and understanding the overall behavior of the models. For example, great performance at the end of the season as we could see for the xG models. On the other hand, the bettor is still mainly interested in the ROI, because it considers the amount of money that was invested into the bets, unlike the cumulative profit/loss.

Chapter 7

Conclusion

In this final chapter, we summarize the outcomes of our research, and their consequences related to the literature mentioned in Chapter 2 and their possible limitations. Firstly, the main goal of the thesis was to investigate if it is possible to be profitable from betting on the outcomes of football matches using the Expected Goals (xG) statistics and also what is the performance compared to models with variables used commonly in the existing literature. Moreover, as the xG statistics is relatively new in football, there has been only one study on the predictive ability of xG variables from Partida *et al.* (2021), where authors suggested using more advanced modeling approaches. Thus, the thesis aimed to extend the study with the introduction of logistic regression and a greater number of variables in the model.

The variables for the model with traditional variables were mainly created based on the study from van Wijk (2021), who conducted very detailed research. However, there were two issues with the variables that he created. Firstly, many variables were correlated, meaning that his models had to suffer from multicollinearity issues, so several variables had to be removed from our model to resolve this issue. Additionally, the author uses datasets that are not available publicly, which might have slightly lowered the accuracy of my model as few variables had to be created with slightly different approaches as explained in Section 4.4.1. The overall accuracy of the model with traditional variables was 65.28%, and thus model served as a sufficient baseline for comparison.

The xG variables were created similarly to the traditional variables based on existing literature (van Wijk (2021) Partida *et al.* (2021)), but from xG values that were gathered from two sources Understat[1] and FootyStats[2] as the xG values differ across models. From these variables, four models were created, two based solely on variables created from xG statistics (one for xG variables created based on Understat xG values and the other based on FootyStats values), and two variables combined from traditional and xG variables. Additionally, the combined models had many insignificant variables, so two more models were created containing only the statistically significant variables.

The dataset used in the thesis included five seasons from 2018 to 2023 across the 5 highest-ranked European leagues and was split into a training set and a testing set. The testing of performance was conducted for Season 4 (2021/2022) and Season 5 (2022/2023), while the logistic regression was fit on data including all previous seasons and game weeks of a particular season as explained in Section 4.3. The first evaluation metric was the accuracy of all models, resulting in the combined model including only statistically significant xG variables based on Understat xG values and traditional variables, achieving the highest accuracy of 65.61% across both seasons.

Lastly, the main objective of this thesis was the betting part, which was divided into two parts based on the strategy. The main goal of betting was to achieve the highest return on investment possible, so the maximal closing odds were used along with the odds from a betting site called B365[3]. For each strategy, several money management techniques were used to achieve the best ROI. The first strategy was betting on the outcome of every match in the season based on the probabilities of our models. The model containing solely Understat xG variables had 4.18% ROI across both seasons using the Kelly Criterion for betting, which was the best result compared to other models. Also, the model from FootyStats xG variables had the second-best ROI, beating both traditional and combined models. The Understat model was the only model that achieved profitable results in both seasons when betting on odds from B365. The second strategy was to recognize mispriced matches based on the difference in probabilities from our models and implied probabilities from the bookmaker's odds, which is the time when an investor can make the most money. In this strategy, the results were better for all models, but this time the combined models were the best performing. Particularly, the combined model from traditional and FootyStats xG variables reached 10.87% ROI across both

[1]understat.com
[2]footystats.org
[3]www.bet365.com

seasons using the Kelly Criterion. Even though the traditional model performed better compared to the models based only on xG variables, the models were able to achieve profitable results using both strategies. The Kelly Criterion had performed the best from all money management techniques across both seasons and both betting strategies, so it was determined to be the best technique for all models.

To summarize, the results of this thesis are positive as the betting results show predictive power for variables created from xG statistics, which supports the claim from Partida *et al.* (2021) that under certain betting conditions, their models based on xG variables were profitable. However, it is important to mention that the research has several limitations and there is a potential for further research. Firstly, the structure of the dataset is limited to only publicly available data compared to the study from van Wijk (2021) as mentioned above. The xG variables in the thesis were created from xG values gathered from two publicly available sources, which might not be the two most accurate models compared to others. This was partly enforced by the year this thesis was written, so further analysis of xG statistics is expected to have better access to additional sources of this metric. Moreover, the Expected Goals (xG) statistics is only the basic building block for other metrics such as Expected Points (xPts) and Expected Assists (xA). The introduction of such variables in the model might considerably increase the accuracy, which directly implies better prerequisites for better results in betting. Thus, this study might serve as a valuable starting point for further analysis of this topic. The researchers might benefit from the knowledge of what variables to include in their models and which betting strategies to use as such a study was not done before.

Bibliography

Bujang, M. A., N. Sa'at, T. M. I. T. A. Bakar, L. C. Joo *et al.* (2018): "Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data." *The Malaysian journal of medical sciences: MJMS* **25(4)**: p. 122.

Buursma, D. (2010): "Predicting sports events from past results." In "14th Twente Student Conference on IT," volume 21.

Caliwag, J. A., M. C. R. Aragon, R. E. Castillo, & E. M. S. Colantes (2018): "Predicting basketball results using cascading algorithm." In "Proceedings of the 1st International Conference on Information Science and Systems," pp. 64–68.

Chang, V., K. Hall, & L. M. T. Doan (2023): "Football results prediction and machine learning techniques." *International Journal of Business and Systems Research* **17**: pp. 565–586.

Constantinou, A. & N. Fenton (2013): "Profiting from arbitrage and odds biases of the european football gambling market." *The Journal of Gambling Business and Economics* **7**: pp. 41–70.

Craney, T. A. & J. G. Surles (2002): "Model-dependent variance inflation factor cutoff values." *Quality Engineering* **14(3)**: pp. 391–403.

Daoud, J. I. (2017): "Multicollinearity and regression analysis." In "Journal of Physics: Conference Series," volume 949, p. 012009. IOP Publishing.

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

Elfrink, T. (2018): "Predicting the outcomes of mlb games with a machine learning approach." Business Analytics Research Paper.

Gambling Commission UK (2022): "Uk sports betting by sport." https://www.businessofapps.com/data/sports-betting-app-market/.

Grant, A. & P. W. Buchen (2012): "A Comparison of Simultaneous Kelly Betting Strategies." *Journal of Gambling Business and Economics* **6(2)**: pp. 1–28.

H2 Gambling Capital (2022): "Europe's gambling market revenue." https://www.egba.eu/eu-market/.

Hsu, Y.-C. (2020): "Using machine learning and candlestick patterns to predict the outcomes of american football games." *Applied Sciences* **10(13)**.

Hubáček, O., G. Šourek, & F. Železný (2019): "Exploiting sports-betting market using machine learning." *International Journal of Forecasting* **35(2)**: pp. 783–796.

Hucaljuk, J. & A. Rakipovic (2011): "Predicting football scores using machine learning techniques." *2011 Proceedings of the 34th International Convention MIPRO* pp. 1623–1627.

Joseph, A., N. Fenton, & M. Neil (2006): "Predicting football results using bayesian nets and other machine learning techniques." *Knowledge-Based Systems* **19(7)**: pp. 544–553. Creative Systems.

Kahn, J. (2003): "Neural network prediction of nfl football games." *World Wide Web Electronic Publication* .

Kassambara, A. (2018): *Machine learning essentials: Practical guide in R.* Sthda.

Kelly Jr., J. L. (1956): "A new interpretation of information rate." *Bell System Technical Journal* **35(4)**: pp. 917–926.

Koseler, K. & M. Stephan (2018): "Machine learning applications in baseball: A systematic literature review." *Applied Artificial Intelligence* **31**: pp. 1–19.

LABRADOR, F. J. & M. VALLEJO-ACHÓN (2020): "Prevalence and characteristics of sports betting in a population of young students in madrid." *Journal of Gambling Studies* **36**: pp. 297–318.

LEUNG, K. (2021): "Assumptions of logistic regression, clearly explained." *Towards Data Science* .

LOEFFELHOLZ, B., E. BEDNAR, & K. BAUER (2009): "Predicting nba games using neural networks." *Journal of Quantitative Analysis in Sports* **5**: pp. 7–7.

MAHER, M. J. (1982): "Modelling association football scores." *Statistica Neerlandica* **36(3)**: pp. 109–118.

MARKOWITZ, H. (1952): "Portfolio selection." *The Journal of Finance* **7(1)**: pp. 77–91.

MAURATH, L. (2020): "Do naive xG models underestimate expected goals for top teams?" `https://www.thesignificantgame.com/portfolio/do-naive-xg-models-underestimate-expected-goals-for-top-teams/`.

NEWALL, P. W. S. (2015): "How bookies make your money." *Judgment and Decision Making* **10(3)**: pp. 225–231.

NGUYEN, Q. H., H.-B. LY, L. S. HO, N. AL-ANSARI, H. V. LE, V. Q. TRAN, I. PRAKASH, & B. T. PHAM (2021): "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil." *Mathematical Problems in Engineering* **2021**: p. 15.

OWRAMIPUR, F., P. ESKANDARIAN, & F. S. MOZNEB (2013): "Football result prediction with bayesian network in spanish league-barcelona team." *International Journal of Computer Theory and Engineering* pp. 812–815.

PARTIDA, A., A. MARTINEZ, C. DURRER, O. GUTIERREZ, & F. POSTA (2021): "Modeling of football match outcomes with expected goals statistic." *Journal of Student Research* **10(1)**.

POLLARD, R. (1985): "69.9 goal-scoring and the negative binomial distribution." *The Mathematical Gazette* **69(447)**: pp. 45–47.

PRASETIO, D. & D. HARLILI (2016): "Predicting football match results with logistic regression." In "2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)," pp. 1–5.

Purucker, M. (1996): "Neural network quarterbacking." *IEEE Potentials* **15(3)**: pp. 9–15.

Rácz, A., D. Bajusz, & K. Héberger (2021): "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification." *Molecules* **26(4)**: p. 1111.

Ricijas, N., D. Dodig Hundric, & A. Huic (2016): "Predictors of adverse gambling related consequences among adolescent boys." *Children and Youth Services Review* **67**: pp. 168–176.

Rue, H. & Ă. Salvesen (2001): "Focus on Sport: Prediction and Retrospective Analysis of Soccer Matches in a League." *Journal of the Royal Statistical Society Series D: The Statistician* **49(3)**: pp. 399–418.

Sauer, R. D. (1998): "The economics of wagering markets." *Journal of Economic Literature* **36(4)**: pp. 2021–2064.

Schreiber-Gregory, D. & K. Bader (2018): "Logistic and linear regression assumptions: Violation recognition and control." *Henry M Jackson Foundation* **247**: p. 22.

Sigler, L. (2003): *Fibonacci's Liber Abaci: a translation into modern English of Leonardo Pisano's book of calculation.* Springer Science & Business Media.

Singh, V., M. Pencina, A. Einstein, J. Liang, D. Berman, & P. Slomka (2021): "Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging." *Scientific Reports* **11(1)**: p. 14490.

Singla, R. & D. A. Singh (2020): "Sports prediction using machine learning." *International Journal of Emerging Technologies and Innovative Research* **7(10)**: pp. 2759–2765. UGC and ISSN Approved.

Smoczynski, P. & D. Tomkins (2010): "An explicit solution to the problem of optimizing the allocations of a bettor's wealth when wagering on horse races." *The Mathematical Scientist* **35**.

Uhrín, M., G. šír, O. Hubáček, & F. Železný (2021): "Optimal sports betting strategies in practice: an experimental review." *IMA Journal of Management Mathematics* **32**.

VEROUTSOS, E. (2023): "The most popular sports in the world." https://www.worldatlas.com/articles.

WHITMORE, J. (2023): "What is expected goals (xG)?" https://theanalyst.com/eu/2023/08/what-is-expected-goals-xg/.

VAN WIJK, D. (2021): *Beating the Bookmakers using Machine Learning.* Master's thesis, Erasmus University Rotterdam, Rotterdam. Erasmus School of Economics.

WOOLDRIDGE, J. M. (2012): *Introduction to Econometrics: A Modern Approach.* South-Western College Pub, 5 edition.

Appendix A

Explanation of Variables

A.1 Traditional Variables - Table 3.9

Season - number of the season in the dataset
GameWeek - number of game weeks in the current season
HomeELO - ELO from the start of the season(rating based on several aspects) of the team playing at home
AwayELO - ELO from the start of the season(rating based on several aspects) of the team playing away
HomePreMatchPosition - position in the league table of the team playing at home before the match
AwayPreMatchPosition - position in the league table of the team playing away before the match
FTR - the result of the match (1 - Home Win, 0 - Draw / Away Win)
B365H - B365 odds for the home team winning
B365DA - B365 odds for the away team not losing
MaxCH - maximum closing odds for the home team from several bookmakers
MaxCDA - maximum closing odds for the away team not losing from several bookmakers
HTAS - home team attacking strength
ATAS - away team attacking strength
HTDW - home team defensive weakness
ATDW - away team defensive weakness
AvgHTP - average points per game for a home team before the match
AvgATP - average points per game for an away team before the match
AvgHTGD - average home team goal difference

AvgATGD - average away team goal difference
HomeForm - a form of home team based on points obtained
AwayForm - a form of away team based on points obtained
HomeTeamH2HRatio - Home team results against the current opponent in the last 5 matches
AwayTeamH2HRatio - Away team results against the current opponent in the last 5 matches
AvgHTSOT - average home team number of shots on target per match
AvgATSOT - average away team number of shots on target per match
AvgHTC - average home team number of corners per match
AvgATC - average away team number of corners per match
AvgHTCAgainst - average number of corners against the home team per match
AvgATCAgainst - average number of corners against the away team per match
AvgHTF - average home team number of fouls per match
AvgATF - average away team number of fouls per match
HomeWinRatioSoFar - the ratio of home wins for a home team when playing at home in the current season
AwayNotLosingRatioSoFar - the ratio of not losing for an away team when playing away in the current season

A.2 xG Variables - Table 3.10

HTxGASFooty - home team attacking strength based on xG from FootyStats
ATxGASFooty - away team attacking strength based on xG from FootyStats
HTxGASUnder - home team attacking strength based on xG from Understat
ATxGASUnder - away team attacking strength based on xG from Understat
HTxGDWFooty - home team defensive weakness based on xG from FootyStats
ATxGDWFooty - away team defensive weakness based on xG from FootyStats
HTxGDWUnder - home team defensive weakness based on xG from Understat
ATxGDWUnder - away team defensive weakness based on xG from Understat
HTAEFooty - home team attacking efficiency based on xG from FootyStats

ATAEFooty - away team attacking efficiency based on xG from FootyStats
HTAEUnder - home team attacking efficiency based on xG from Understat
ATAEUnder - away team attacking efficiency based on xG from Understat
HTDEFooty - home team defensive efficiency based on xG from FootyStats
ATDEFooty - away team defensive efficiency based on xG From FootyStats
HTDEUnder - home team defensive efficiency based on xG from Understat
ATDEUnder - away team defensive efficiency based on xG from Understat
HomexGShotFooty - home team average xG per shot in the current season based on xG from FootyStats
AwayxGShotFooty - away team average xG per shot in the current season based on xG from FootyStats
HomexGShotUnder - home team average xG per shot in the current season based on xG from Understat
AwayxGShotUnder - away team average xG per shot in the current season based on xG from Understat
HomeFootyForm - a form variable of a home team in the last 5 matches based on xG from FootyStats
AwayFootyForm - a form variable of an away team in the last 5 matches based on xG from FootyStats
HomeUnderForm - a form variable of a home team in the last 5 matches based on xG from Under
HomeUnderForm - a form variable of an away team in the last 5 matches based on xG from Under
HomeFootyPos - home team league position based on xGD in the current season based on FootyStats xG
AwayFootyPos - away team league position based on xGD in the current season based on FootyStats xG
HomeUnderPos - home team league position based on xGD in the current season based on Understat xG
AwayUnderPos - away team league position based on xGD in the current season based on Understat xG

Appendix B

Development of Coefficients

Beta Coefficients for ModelTraditional - Season 4

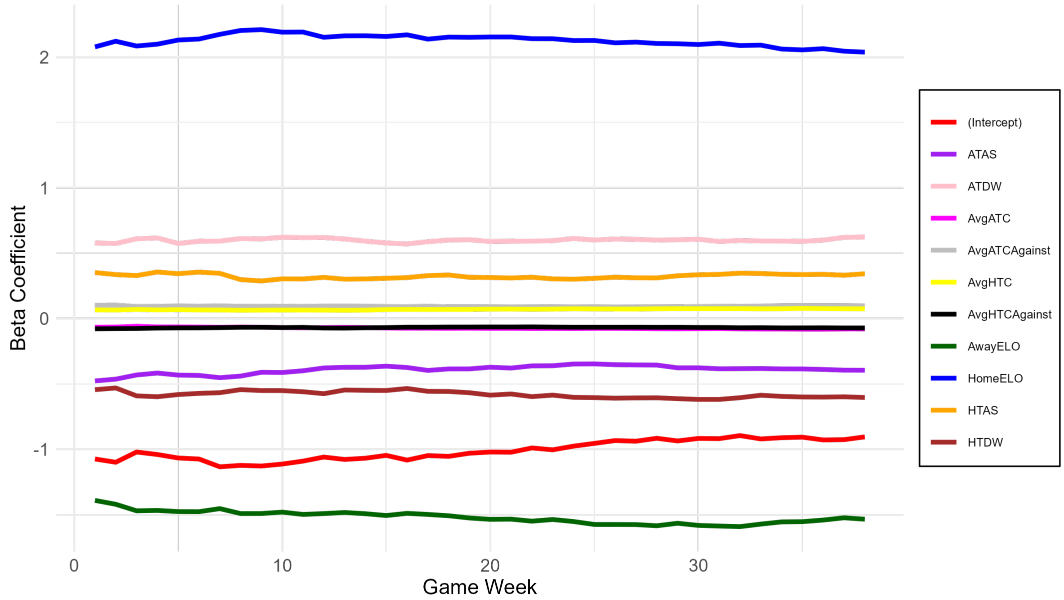

Figure B.1: Traditional Model Season 4

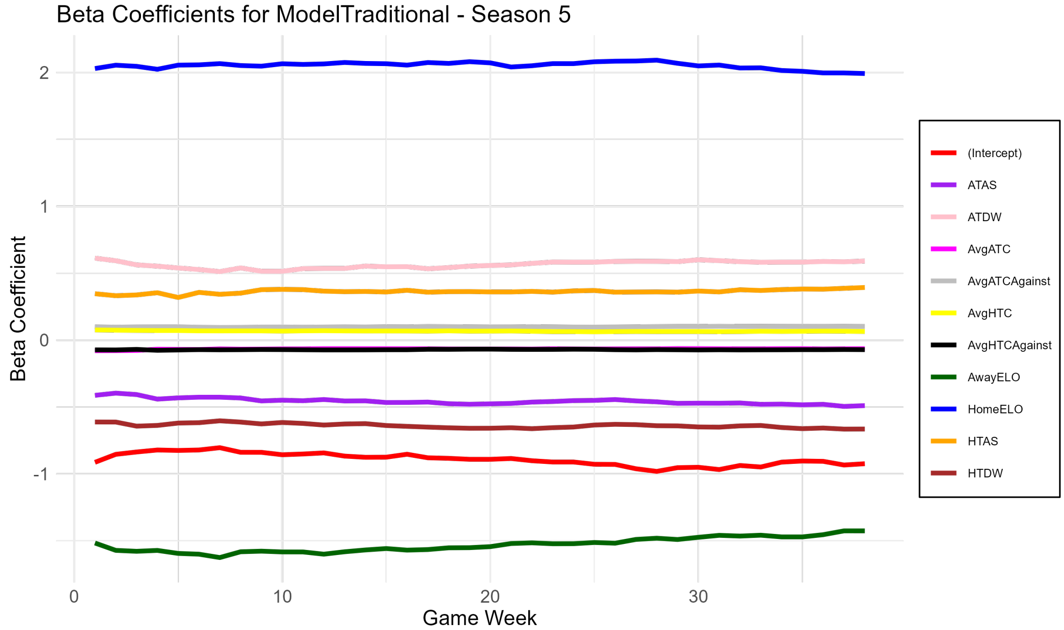

Figure B.2: Traditional Model Season 5

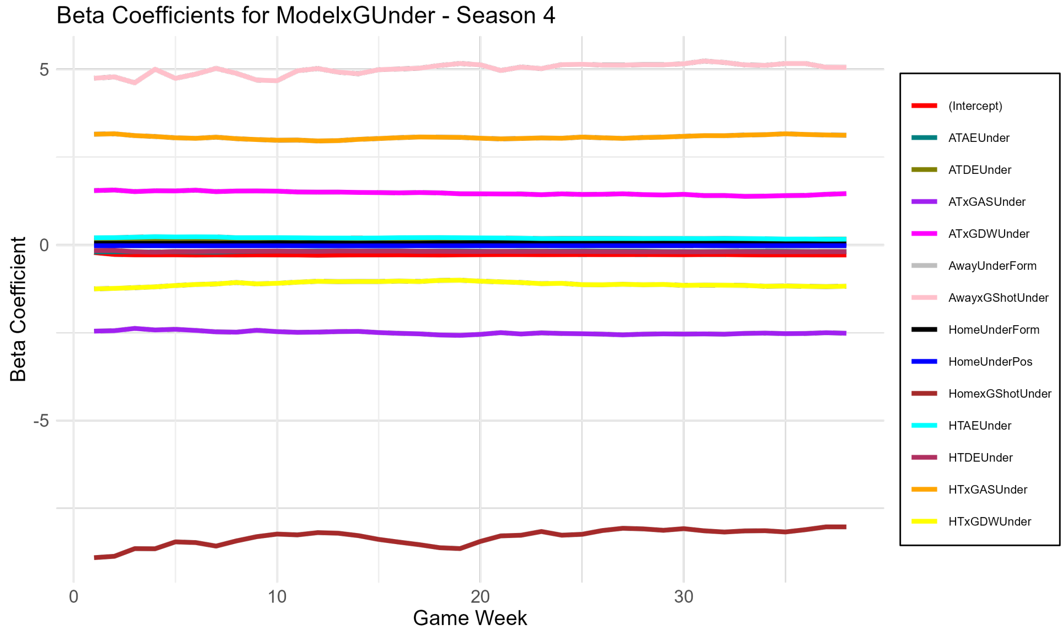

Figure B.3: Understat xG Model Season 4

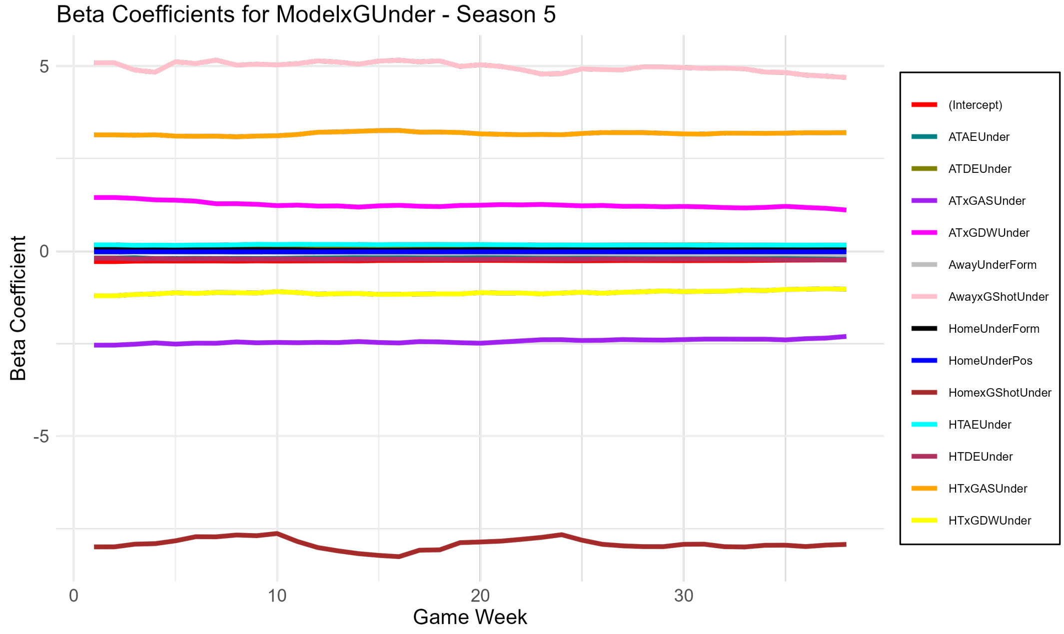

Figure B.4: Understat xG Model Season 5

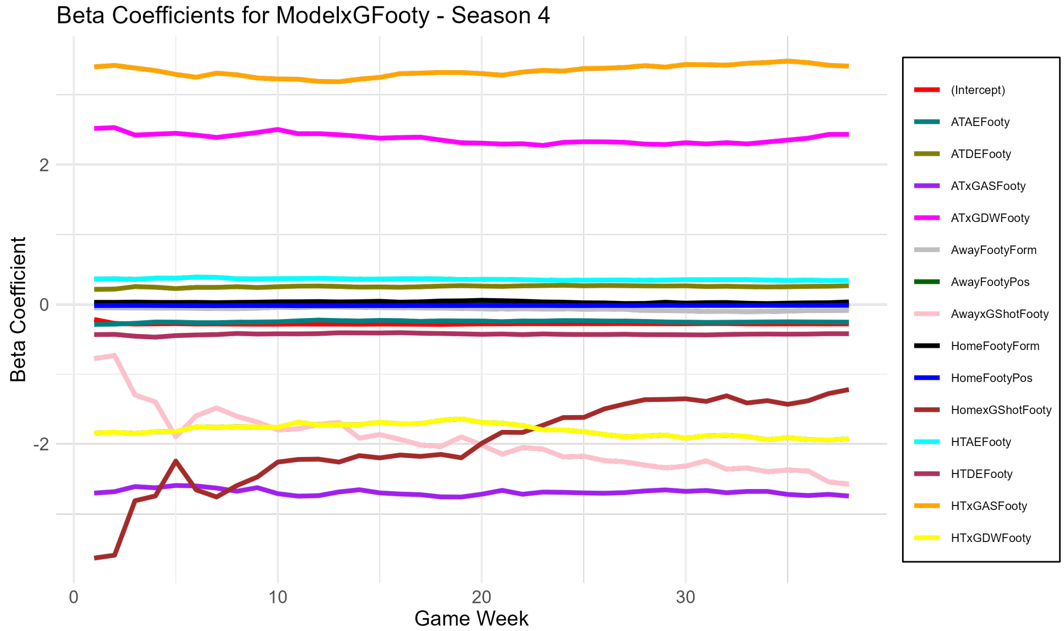

Figure B.5: FootyStats xG Model Season 4

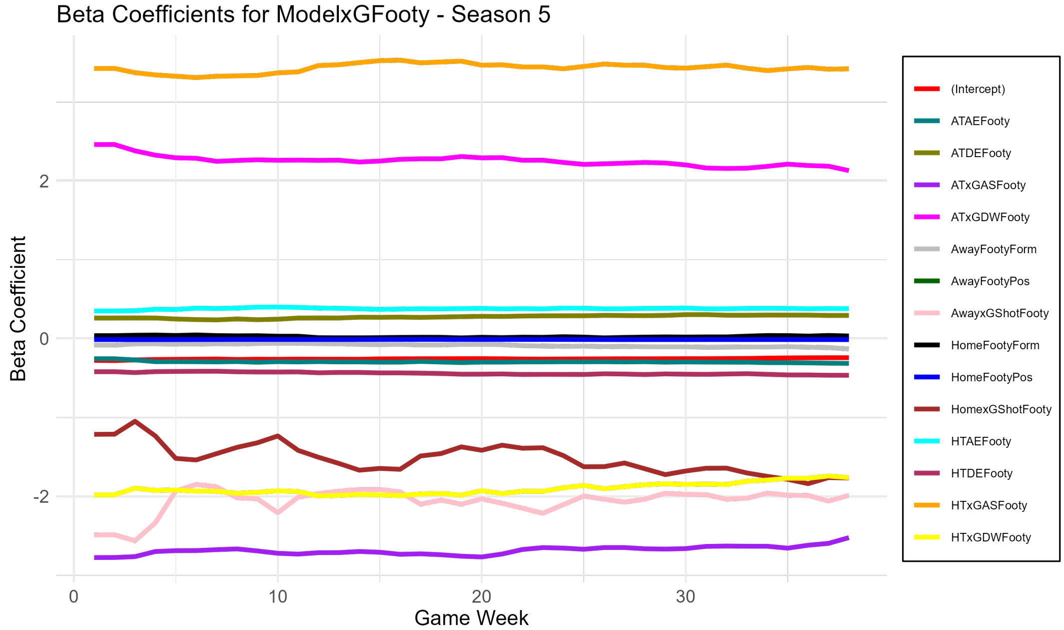

Figure B.6: FootyStats xG Model Season 5

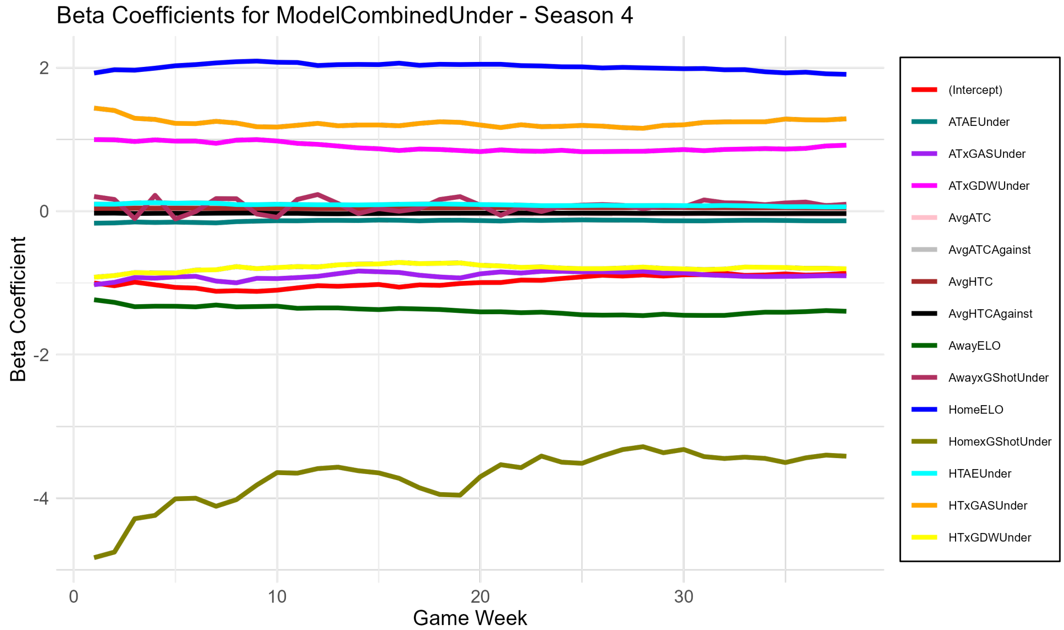

Figure B.7: Combined Understat Model Season 4

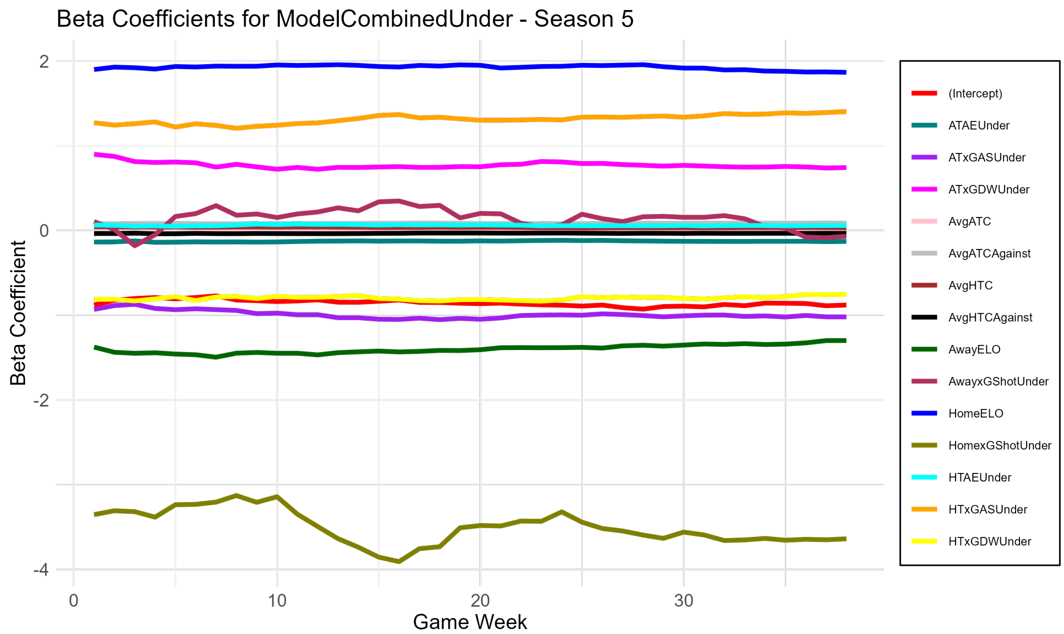

Figure B.8: Combined Understat Model Season 5

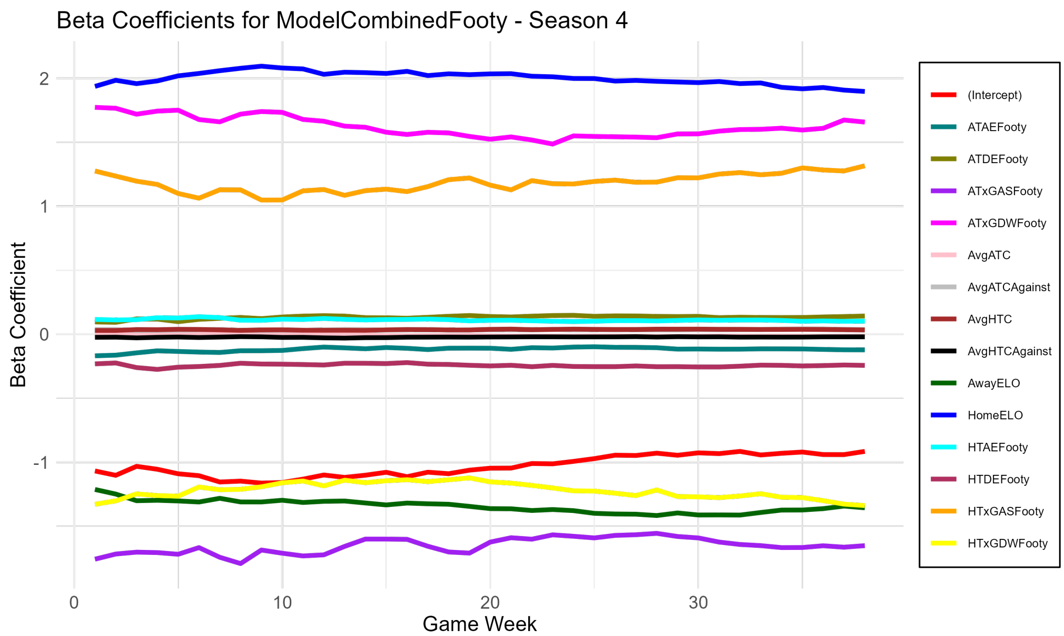

Figure B.9: Combined FootyStats Model Season 4

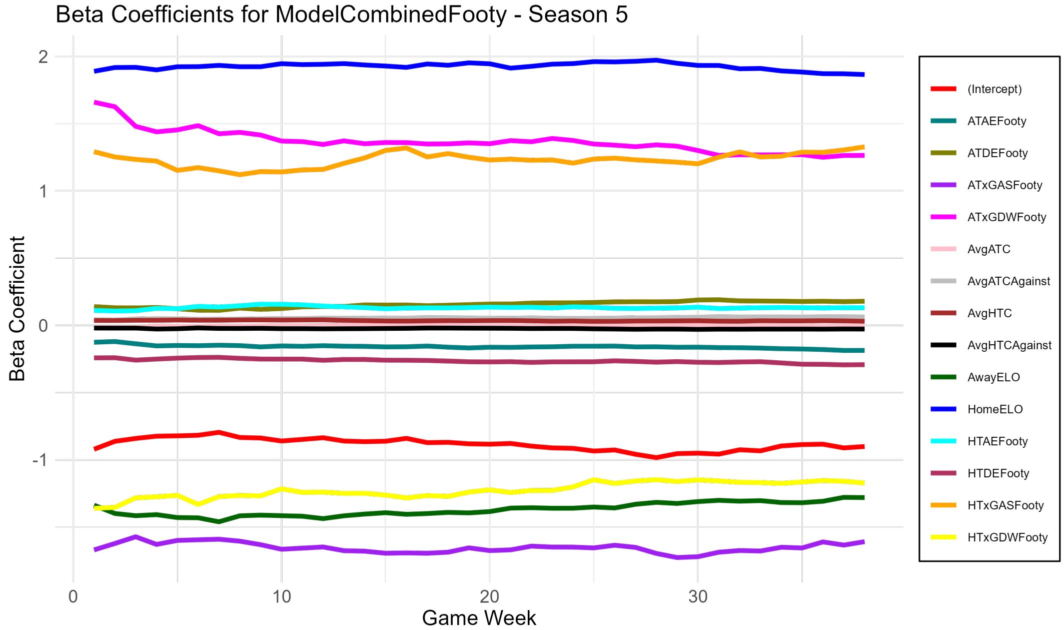

Figure B.10: Combined FootyStats Model Season 5