# Assessing the Accuracy of Betting Odds in European Football

Joseph Edward Pym
8404110

*Supervisors:*:
Dr. Maha Moustafa *and* Dr. Alun Owen

*A dissertation presented in the School of Computing, Electronics & Mathematics, Coventry University, for the degree of Bachelor of Science in Mathematics & Statistics.*

Submitted April 2021

## 0.1 Abstract

This dissertation aims to quantify bookmaker accuracy across three different markets: the 1X2, Under/Over 2.5 Goals, and Asian Handicap, using data from a range of *elite* European football leagues and different tiers in the English & Scottish football league pyramids, from the 2005/06 until the 2019/20 seasons. Using a wide range of statistical tools such as correlation analysis and principal components analysis, measures of accuracy are explored including the coefficient of determination $R^2$, the root-mean-square-error RMSE, and values of predictive accuracy $P_1$ and $P_2$, using the findings to create a simple accessible algorithm for placing bets. Bookmaker accuracy is found to be extremely high in the 1X2 Home Win and Away Win, and Asian Handicap markets, but poor in the 1X2 Draw and Under/Over 2.5 Goals markets, though the latter is improving over time. In addition, competitive balance is found to be a factor affecting accuracy—with more competitive leagues (such as Germany's Bundesliga and France's Ligue Une) having worse bookmaker performance than uncompetitive leagues (Portugal's Primiera Liga). Bookmaker commissions are both higher in lower levels and reducing over time in the 1X2 and Under/Over 2.5 Goals markets, whilst remaining constant in the Asian Handicap markets. Using these findings, an algorithm based on accepting high bookmaker accuracy in the 1X2 Home Win and Away Win and Asian Handicap markets is proposed, which—despite winning 50.5% of bets—makes a loss of 4.3% on 71,144 bets. This is compared against a random bet strategy, which despite losing 11.8% more bets, only makes a further loss of 6.7%, indicating that accuracy in placing bets does not directly equate to profitability or winnings.

## *Acknowledgements*

I would like to take this opportunity to give my thanks:

- to my supervisors, Maha and Alun, for their help and expertise throughout the year. I couldn't have asked for any more of them;
- to Claudia, for her constant love and support;
- and to my family, for only being a phone call away whenever needed.

Thank you.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

## 0.2 Disclaimer

Involvement in gambling can often lead to highly dangerous and damaging consequences, both for the bettor and those around them: *problem gambling* is characterised as 'persistent and recurrent problematic gambling behaviour leading to clinically significant impairment or distress' (American Psychiatric Association, 2018). This dissertation will not seek to solve the social, economic, or political issues surrounding gambling, but will instead focus wholly on the mathematics and statistics upon which sports gambling is based.

# Chapter 1

# Introduction

## 1.1 Background Information

### 1.1.1 Football in Europe

Association football (or soccer) is the most popular sport in the world (Giulianotti, 2012), with over 240 million players worldwide (FIFA.com, 2001). Professional football in Europe is governed by UEFA (the Union of European Football Associations) which each nation having their own governing body (E.G., the Deutscher Fußball-Bund in Germany) administering football in that country (UEFA.com, n.d.[b]), including the league systems, known as pyramids. Generally, the pyramids follow a similar structure, with a number of promotion and relegation places contested for throughout the course of a year-long season, which ensures clubs in each division have a similar level of ability. The top division (or tier) in each country is allocated a number of qualification places to the two Europe-wide club competitions: the Champions' League and Europa League, allowing the best teams in Europe to compete against one other.

The number of such allocated places is chosen via the UEFA country coefficient, which ranks the countries by the performance of their collective clubs in these competitions. The top six[1] of these are chosen to be considered the *elite* European leagues. These are given in Table 1.1. Also considered are the English & Scottish football pyramids given in Table 1.2. These pyramids are considered as `football-data.co.uk` (the data source, discussed in Section 1.3.2) contains data on five English leagues and four Scottish leagues: No other country has more than two leagues worth of data, allowing for a comparison between different tiers in the pyramid.

Table 1.1: The UEFA Country Coefficient for the top six European leagues (UEFA.com, n.d.[a]).

| *Country* | *Top Division* | *Coefficient* |
|---|---|---|
| Spain | La Liga | 92.283 |
| England | Premier League | 90.712 |
| Italy | Serie A | 72.295 |
| Germany | Bundesliga | 71.856 |
| France | Ligue Une | 54.915 |
| Portugal | Premiera Liga | 47.349 |

[1] As of 5th January, 2021.

Table 1.2: The English & Scottish football league pyramids (English Football League, n.d. Scottish Professional Football League, 2021).

| *Tier in Pyramid* | *English Pyramid* | *Scottish Pyramid* |
|---|---|---|
| 1 | Premier League (EPL) | Premier League (SPL) |
| 2 | Championship | Championship |
| 3 | League One | League One |
| 4 | League Two | League Two |
| 5 | Conference/National League | — |

### 1.1.2 Probabilities, Odds, and Gambling

Every event $i$ has a **probability** of occurring, denoted as $\mathbb{P}(\text{Event } i) = p_i$, between 0 (almost never occurs) and 1 (almost certain to occur), with the sum of all possible outcomes being equal to 1 (Grinstead and Snell, 2012). For example, a table of probabilities for rolling a number $X$ between one and six on a fair, six-sided die can be constructed, as follows:

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\mathbb{P}(X = x) = p_i$ | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

In gambling, BOOKMAKERS[2] supply BETTORS with the **odds** of event, rather than the probability of it occurring (Štrumbelj, 2014). In statistics, 'odds' are used in concepts such as odds ratios and risk analysis, with a broad range of applications, I.E., medicine (Morris and Gardner, 1988). Such 'odds' consider the odds *for* an event (how likely it is to occur) and will be referred to as 'statistical odds', which differ from the 'odds' offered by bookmakers, referred to as 'bookmaker odds', which generally consider the odds *against* an event (how unlikely it is to occur). In this section, four types of expressing odds are explained, with an example for each provided.

The statistical odds of an event $i$ with a probability $p_i$ of occurring have the odds of occurring defined in Equation 1.1. From these odds, the corresponding probability can be found using the formula $p_i = (O_i)(O_i + 1)^{-1}$.

$$\text{Statistical, } O_i = \frac{p_i}{1 - p_i} \tag{1.1}$$

Bookmaker odds are displayed in three main methods. Fractional (also known as British) odds, Equation 1.2, are the inverse of the statistical odds. The winnings are computed using the formula Winnings = Stake × Odds + Stake: a £10 bet on a 5/1 winning bet would result in a profit of £50, and a winnings (including the STAKE) of £60.

$$\text{Fractional, } O_i = \frac{1 - p_i}{p_i} \tag{1.2}$$

Decimal (European or Continental) odds, Equation 1.3, are the style chosen by `football-data.co.uk` (Section 1.3.2), and due to this, are the style chosen throughout this project. They represent the inverse of the probability of an event, and thus have a minimum of 1.0 (as the maximum probability is 1). This method includes the stake, thus Winnings = Stake × Odds.

[2]Words in SMALL CAPS are defined in Appendix A.

$$\text{Decimal, } O_i = \frac{1}{p_i} \tag{1.3}$$

*Moneyline* (American) odds, Equation 1.4, can be either negative or positive. If they are negative, they show the stake required to win \$100; if they are positive, they show the winnings from a \$100 bet[3] (Cronin, 2019).

$$\textit{Moneyline, } O_i = \begin{cases} +100 \times \frac{1-p_i}{p_i} & \text{if } p_i \leq 0.5 \\ -100 \times \frac{p_i}{1-p_i} & \text{if } p_i > 0.5 \end{cases} \tag{1.4}$$

---

**Example 1: Computing Odds**

Considering the fair die scenario, the odds of rolling a six are computed, and presented in each style discussed.

Statistical

$$O_6 = \frac{p_6}{1-p_6} = \frac{1/6}{1-1/6} = \frac{1/6}{5/6} = \frac{1}{5}$$

Fractional

$$O_6 = \frac{1-p_6}{p_6} = \frac{1-1/6}{1/6} = \frac{5/6}{1/6} = \frac{5}{1}$$

Decimal

$$O_6 = \frac{1}{p_6} = \frac{1}{1/6} = 6.0$$

*Moneyline*

$$O_6 = +100 \times \frac{1-p_6}{p_6} = +100 \times \frac{1-1/6}{1/6} = +500$$

---

*Consensus odds* refer to the market average (mean) odds for that outcome, for example, if five bookmakers, say A, B, C, D, and E, offered (decimal) odds on an outcome of 5.0, 5.5, 5.3, 5.4, and 4.1 on a Home Win, the consensus odds are found to be 5.1. If odds from only one bookmaker are used instead, the data could go against market beliefs if that bookmaker was an outlier (such as bookmaker E).

Equations 1.1 to 1.4 can be rearranged to make $p_i$ the subject, called the *underlying probability*, denoted $\mathbb{P}_{\text{und}}$, from the odds offered. The offered odds are manipulated to allow for a rate of commission called the OVERROUND: the underlying probabilities will not sum to one (Cortis, 2015; Štrumbelj, 2014).

[3] Or, of course, 100 of whichever currency is being used.

The overround is countered by normalising the underlying probabilities to find the *consensus probabilities*, $\mathbb{P}_{\text{cons}}$. The equation for normalisation is given in Equation 1.5, where the $j^{\text{th}}$ outcome is being normalised; the summation is for all $i$ out of $k$ possible outcomes.

$$\mathbb{P}_{\text{cons}}(\text{Outcome}_j) = \frac{\mathbb{P}_{\text{und}}(\text{Outcome}_j)}{\sum_{i=1}^{k} \mathbb{P}_{\text{und}}(\text{Outcome}_i)} \tag{1.5}$$

A point worth considering: whilst this counters the overall overround, normalisation assumes the bookmaker overround is proportionate across each market (Angelini and De Angelis, 2019), however, this is unlikely, as bookmakers set odds to exploit the bettor. Consensus probabilities may differ slightly with the bookmaker's genuine beliefs, though this is the best method to counter overround (Levitt, 2004).

---

**Example 2: Finding Consensus Probabilities**

To demonstrate this point, a hypothetical example is considered. In a match between Hull City (the Home team) and Arsenal (the Away team), the latter are known to be strong favourites. Assuming the consensus odds (displayed in both the statistical and fractional styles) offered on the result are:

| | *Home Win* | *Draw* | *Away Win* |
|---|---|---|---|
| *Statistical Odds* | 1/5 | 1/1 | 2/1 |
| *Fractional Odds* | 5/1 | 1/1 ('Evens') | 1/2 |

These are converted to the decimal (European) style by adding one to the fractional odds:

| | *Home Win* | *Draw* | *Away Win* |
|---|---|---|---|
| *Decimal Odds* | 6.00 | 2.00 | 1.50 |

The inverse of the decimal odds is taken to find the underlying probabilities. The sum is found, and each underlying probability is divided by this sum, to give the consensus probabilities.

| | *Home Win* | *Draw* | *Away Win* | *Sum* |
|---|---|---|---|---|
| *Underlying Probability* | 0.1667 | 0.5000 | 0.6667 | 1.33 |
| *Consensus Probability* | 0.1250 | 0.3750 | 0.5000 | 1.00 |

This indicates that the bookmakers have predicted Hull City have a 12.5% probability of winning; Arsenal have a 50% probability of winning; and there is 37.5% probability of a draw. The bookmakers have assigned a 33% overround to this match.

---

### 1.1.3 Betting Markets of Interest

There are a vast range of markets used in football betting (bet365, n.d.); the markets of interest in this project are:

- **1X2** (or full-time result)—betting on the explicit final outcome of the match being either a Home Win, a Draw, or an Away Win.
- **Goal Markets**—this is a bet on the amount of total goals or number of goals for either side, usually as half-numbers (E.G., 0.5, 1.5, 2.5, ETC.), with odds offered for Under/Over the given amount. Throughout this project, the Under/Over 2.5 Goals (UO) market, considering total goals in the match, is investigated.
- **Asian Handicap**—this style of betting allows a seemingly one-sided fixture to become competitive. For Asian Handicap (AH) bets, a handicap given can be a whole number or half-number, used as a head-start (if positive) or a detriment (if negative) to a team; for example, if, in a hypothetical match between Hull City and Arsenal, Hull City's handicap was +2.5, Arsenal would need to win by 3 goals in order for a bet on them to win: Hull City would just need to avoid a loss by 3 goals for a bet on them to win. In addition, Asian handicaps can also be quarter-numbers (such as $\frac{3}{4}$): the bet is then split into two: half the stake on nearest half-number and half on the nearest integer: $\frac{1}{2}$ and half on 1, in this case.

Gambling is a huge part of football culture, with 27 of the 44 teams (61.3%) in the English Premier League and Championship having a gambling company as their main shirt sponsor (Davey, 2020); those that don't will often have a betting company as a 'Club Partner.' For example, Arsenal (who have Fly Emirates as their shirt sponsor) are partnered with SportsBet.io; and Manchester City (Etihad Airways) are partnered with Marathon Bet (The Football Association Premier League Limited, 2019). The combined income of betting partnerships in the Premier League is around £70 million. Eight of the 20 Spanish La Liga sides have a gambling company on their shirts: BetWay alone sponsor three (Score and Change, 2020). Clearly, bookmakers are highly invested in football teams. This leads one to wonder: 'How accurate are the betting odds offered by bookmakers in European football matches?'

## 1.2 Literature Review

### Assessing Accuracy

Whilst there is literature into gambling, surprisingly, none explicitly aims to investigate and to quantify the accuracy of the odds offered by bookmakers in depth. Some papers look briefly into the accuracy to develop further ideas, such as Kaunitz, Zhong, and Kreiner (2017), which found coefficient of determination $R^2$ of 0.999, 0.995, and 0.998 for Home Wins, Draws and Away Wins respectively in the 1X2 market, showing extremely high accuracy. The paper, however, did not probe further into the accuracy of the odds.

The concept of *efficiency* in betting markets has been investigated by several papers to find areas where bookmakers can be exploited. Kuypers (2000) investigated the efficiency in English football assuming a fixed overround of 11% (based on the average overround in a small sample, $n = 3382$), and found that bookmakers can maximise profits by taking advantage of bookmaker behaviour (I.E., team loyalty); the paper, however, only considered two seasons (1993/94 and 1994/95) and only in the 1X2 market. This assumption is investigated in Section 3.5, across more leagues, a

longer period, and in different markets. Bookmaker efficiency was also investigated by Angelini and De Angelis (2019) who considered a much wider dataset—41 bookmakers on 11 European championships over 11 seasons—and found that the efficiency varied across bookmakers, and that—when considering the best odds—four leagues showed inefficiencies in the 1X2 market. Štrumbelj and Šikonja (2010) showed a discrepancy between the forecasts and bookmakers, and that some leagues have 'significantly better forecasts' than others, with the French Ligue Une having the 'most evenly matched' teams, and that bookmaker's forecasts are increasing in efficiency over time; Cain et al. (2000) found evidence of the FAVOURITE-LONGSHOT BIAS in one bookmaker (William Hill) across 2855 English football matches in the 1991/92 season, where favourites will win more than implied by market probabilities, and longshots will win less. It was also found, using a Chi-Squared $\chi^2$ test, that the number of home and away goals are independent.

COMPETITIVE BALANCE is investigated in a number of papers. Some propose methods of quantifying it in football (Goossens, 2005; Ramchandani, 2012). Others consider its implications in other sports (I.E., Rottenberg (1956) considers baseball). A link between competitive balance and the accuracy of bookmakers in football is a gap in present literature.

Khazaal et al. (2012) investigated whether 'knowledge and expertise on football led to better prediction skills for match outcomes' and found that—in a study of 258 participants (21.3% 'experts', 24.4% 'amateurs', 54.3% 'laypersons')—expertise did not affect the accuracy, showing the 'belief that football expertise improves betting skills is no more than a cognitive distortion.' Therefore, in Chapter 4, an ALGORITHM is constructed that removes all subjectivity, instead focussing wholly on the odds involved.

## A Simple Betting Method

There is no published method for bettors that involves information from one, or a limited number of, bookmaker(s) that is simple and repeatable.

Various techniques and DISTRIBUTIONS have been applied to data to predict the final score of a football match: Dixon and Coles (1997) applied a Poisson Regression Model, using Maximum Likelihood Estimates to exploit market inefficiencies to decide where to bet; Owen (2009) used a Possion Dynamic Generalised Linear Model with parameters derived from Markov-Chain-Monte-Carlo (MCMC) methods in a Bayesian 'framework'; Karlis and Ntzoufras (2009) similarly use MCMC methods, but instead used Skellam's distribution to model goal difference; and Constantinou (2020) investigated the efficiency of the Asian Handicap market in relation to the 1X2 market, using a Hybrid Bayesian Network to simulate relationships between varying statistics, such as possession, shots, and goals.

There also exist methods that require large amounts of data, such as the method proposed by Kaunitz, Zhong, and Kreiner (2017), which is based on comparing the 'maximum' odds offered for an outcome to the consensus odds offered from a large number of bookmakers for that same outcome and placing a bet if the maximum is sufficiently higher, or large amounts of data used alongside high-level tools, such as one proposed by Godin et al. (2014), which used pattern recognition in the form of text processing Twitter posts to pool opinion by users as a form of expert elicitation (50 million such posts were used when betting on half a football season). The algorithm produced in this project, therefore, should require lower amounts of input data without advanced tools, thus being more accessible to bettors.

## 1.3 Rationale, Aims, Objectives and Methods

This project aims to fill a gap in literature: no paper investigates the accuracy of bookmakers in-depth, across a large dataset; nor does one provide a simple, accessible algorithm for bettors to follow and be able to objectively place bets.

The aim of this project to assess the accuracy of betting odds offered by a range of bookmakers in European football. It aims to look into different markets—the 1X2, Under/Over 2.5 Goals (UO), and Asian Handicap (AH) markets—and will aim to determine factors that may or may not influence the accuracy of the odds, such as the level, season, or country (league) of the match, and will also look briefly into the overround, where the odds are lowered, allowing bookmakers to make profit.

In order to investigate this, data from `football-data.co.uk`—a website set up by Joseph Buchdahl, a betting analyst and author of multiple published works about betting (Buchdahl, n.d.[b])—is utilised, which provides historical results and odds in easy-to-access comma-separated-value (`.csv`) files (Buchdahl, n.d.[c]). With this data, different angles for assessing the accuracy of the odds are explored, including exploratory analysis (Hoaglin, 1977) and calculating the predictive power of the odds (Owen, 2009). In addition, a range of visual aids such as tile plots, histograms, and density plots are used, and linear models are created to predict the actual probability of an event from the bookmaker consensus probability.

Throughout the project, the programming language R and a range of additional packages (Section 1.3.1) are used, allowing for the tasks to be performed efficiently (R Core Team, 2021).

The objectives for the project are as follows: first, the accuracy of odds offered in the 1X2 market in *elite* European leagues are to be assessed. This analysis will be combined with a review into whether or not competitive balance impacts the accuracy of the odds offered. Secondly, accuracy of odds offered in different levels of the English & Scottish league pyramids are to be assessed, looking at whether the accuracy of odds differs across levels of football, looking into the three aforementioned markets. This is combined this with an investigation into the overround. Finally, an algorithm to place bets, using previous findings to assist in the development, will be constructed, and tested against a random bet strategy: placing a similar number of bets at random.

Upon the completion of these objectives, the findings will be presented in a conclusion, with a discussion into the implications and areas for future research.

### 1.3.1 R Packages Used

- `car` — 'Companion to Applied Regression' (Fox and Weisberg, 2019).
- `MASS` — Support for statistical functions and tests, such as the $\chi^2$ test of independence (Venables and Ripley, 2002).
- `ggplot2` — For elegant graphics and a wide range of plots and graphs (Wickham, 2016).
- `gridExtra` — To allow for multiple-figure plots created by `ggplot2` (Auguie, 2017).
- `scales` — 'Scale functions for visualisation' (Wickham and Seidel, 2020).
- `e1071` — Miscellaneous statistical functions, used in this project for the `discrete` distribution tools (Meyer et al., 2020).

### 1.3.2 The Data

#### Sourcing the data

The data is collected from `football-data.co.uk` (F-D). F-D use the following bookmakers for their consensus odds (as of 02/10/20): Bet365, Blue Square Bet, Bet & Win, Gamebookers, Interwetten, Ladbrokes, Pinnacle, Sporting Odds, Stan James, BetVictor, and William Hill, as well collecting odds from BetBrain (an odds comparison website, using information from up to 80 bookmakers per match). Betting odds are taken from the individual bookmakers and compiled. The odds for matches during the weekend are collected on Friday afternoons; odds for midweek matches are collected on Tuesday afternoons. For game statistics, such as home/away corners, free kicks, shots (on/off target), offsides and cards, F-D uses BBC Sport, ESPN Soccer, `Gazzetta.it` and `Football.fr` (Buchdahl, n.d.[a]).

Throughout the project, data from the leagues in Tables 1.1 and 1.2 is considered from the 2005/06 season until the 2019/20 seasons. This is due to `football-data.co.uk` having sufficient data on all leagues in these seasons. The columns from the F-D `.csv` files that are required for the analyses are given in Table 1.3.

Table 1.3: Columns from `football-data.co.uk`'s datasets used in this project.

| *Name* | *Meaning* |
|---|---|
| `div` | Division. |
| `date` | Date of the match. |
| `HomeTeam`, `AwayTeam` | The teams involved in the match. |
| `FTHG`, `FTAG` | Full-time home/away goals. |
| `FTR` | Full-time result: Equal to 'H' for a Home Win, 'A' for an Away Win, and 'D' for a Draw. |
| `BbAvH`, `BbAvA`, `BbAvD` | The bookmaker consensus odds for a Home Win, Away Win and Draw, respectively. Renamed `AvgH`, `AvgA` and `AvgD` from 19/20. |
| `BbAv>2.5`, `BbAv<2.5` | (Rendered as `BbAv.2.5` and `BbAv.2.5.1` in R.) The bookmaker consensus odds for Over 2.5 Goals and Under 2.5 Goals. Renamed `Avg>2.5` and `Avg<2.5` from 19/20. |
| `BbAvAHh` | The bookmaker consensus home handicap offered. Renamed `AHh` from 19/20. |
| `BbAvAHH`, `BbAvAHA` | The bookmaker consensus odds for the Asian Handicap market. Renamed `AvgAHH` and `AvgAHA` from 19/20. |

#### Data Storage

With R, the `.csv` files from F-D can be accessed without downloading them, eliminating the need for local data storage, using the URL in place of a file name with the `read.csv("")` command. Further, a `for` loop is used to download and store all the datasets in R; this is shown in Section 2.1. After reading in the data, each league is given a 'country code':

| Germany | England | Spain | France | Italy | Portugal | Scotland |
|---|---|---|---|---|---|---|
| `de` | `en` | `es` | `fr` | `it` | `po` | `sc` |

## 1.4 Structure

Following this introduction, the main body of the dissertation[4] is split into three main chapters: Chapters 2, 3 and 4:

- In Chapter 2, the accuracy of betting odds in *elite* European leagues (defined in Table 1.1) in the 1X2 market is assessed. This is begun by conducting initial (Section 2.1) and exploratory (Section 2.2) data analysis. These steps help assess the suitability of the data, find general trends, and locate areas for further analysis. After this, correlation analysis is conducted (Section 2.9) where linear models are created with a view to predict the actual probability of an event, based on the bookmaker consensus probability. In Sections 2.4, 2.5, and 2.7, statistics found in the correlation analysis, as well as predictive power statistics $P_1$ and $P_2$ are used to compare the accuracy across leagues and seasons, investigating possible reasons behind this. In particular, competitive balance across the *elite* leagues is investigated in Section 2.6, and principal components analysis is conducted in Section 2.8. Finally, the chapter is concluded in Section 2.9.
- In Chapter 3, the English & Scottish league pyramids are considered. Similar analyses on the 1X2 market are conducted, as well as on the Under/Over 2.5 Goals and Asian Handicap markets. This begins with exploratory data analysis (Section 3.1), where basic calculations are found and visual analysis is used to find trends and areas for further study. This is followed by carrying out correlation analysis (Section 3.2) on the data, finding differences and similarities in bookmaker accuracy across different levels of ability, across different markets. Comparisons are made between the different markets, finding where bookmakers perform well, and where they struggle. The chapter is concluded after a brief look into the bookmaker overround over time and across each market (Section 3.5), investigating where bookmakers apply the largest percentage commission.
- In Chapter 4, findings from Chapters 2 and 3 are used to construct an applicable, simple, and repeatable algorithm for bettors to place bets, with the aim to create long-term gain on an investment. First, the algorithm is outlined, which is based on winning as many bets as possible. It is ran on both the *elite* and English & Scottish datasets, on the high-accuracy markets. The performance is reviewed and compared to a random bet strategy in Section 4.4, which is ran multiple times before concluding.

After these chapters, final remarks are given in Chapter 5. The project's findings are discussed, as well as how they can be used and areas for future research are outlined in Section 5.1, and, as the researcher, I discuss what I have learnt personally throughout the course of the project in Section 5.2.

### Appendices

Throughout the dissertation, several words are written in SMALL CAPS: these are defined in Appendix A, with mathematical terms in Appendix A.1 and gambling terms in Appendix A.2. Located in Appendices B, C, and D is the full R code used for Chapters 2, 3, and 4 respectively. Appendix E is the output of the random bet strategy, explained in detail in Section 4.4. The project diary is in Appendix F; the dissertation word count is in Appendix G; and the ethical approval certificate is in Appendix H.

[4]The term 'dissertation' refers to this document (the write-up); 'project' refers to the entire work, such as steps made prior: conducting analyses, computations, ETC.

# Chapter 2

# Assessing the accuracy of betting odds in *elite* European football leagues, from 2005 to 2020.

This chapter will aim to assess the accuracy of betting odds offered by bookmakers in *elite* European leagues, as defined previously (Section 1.1.1, Table 1.1). To begin the analysis—after reading in and CLEANING the data—the techniques of initial and exploratory data analysis are applied, with the former looking at a randomly chosen subset (one league and one season) of data, ensuring it behaves as expected, and the latter looking at four 'major themes' (Hoaglin, 1977) to answer questions about the whole dataset. Visual aids are also created to further explore and analyse the dataset.

Then—using correlation analysis—linear models are created to predict the observed probability of a result from the bookmaker's offered odds: from these models, statistics to test the FIT of the data, such as the root-mean-square-error are found. To aid this, statistics of predictive performance are also used, before looking at the impact of competitive balance on the accuracy of the bookmaker and presenting conclusions. Principal components analysis is conducted as a further method of analysing the impact of the league and season on the bookmaker accuracy.

## 2.1 Initial Data Analysis

Initial data analysis (IDA) is used to answer four questions about the data:

- What is the quality of the data?
- What is the quality of the measurements?
- Did the implementation of the study fulfil the intentions of the research design?
- What are the characteristics of the data sample? (Adèr, 2008)

To answer these questions, simple tasks are conducted on a subset of the full dataset. Chosen at random, the **French Ligue Une 2016/17** season is considered. The tasks to be carried out include calculating the mean consensus probabilities and comparing them with the observed probabilities, and plotting a histogram to assess the distribution of the consensus probabilities.

The first step is to read in the data from `football-data.co.uk`, before removing the columns that are not required. The R code used to complete this step is below.

```
fr_l1_1617 <- read.csv("https://www.football-data.co.uk/mmz4281/1617/F1.csv")
fr_l1_1617 <- fr_l1_1617[,c("Div", "Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG", "FTR",
    "BbAvH", "BbAvD", "BbAvA")]
fr_l1_1617 <- na.omit(fr_l1_1617)
```

As mentioned in Section 1.3, consensus probabilities will be considered, rather than odds (see Section 1.1.2, Equation 1.5). The R code used to conduct this is below.

```
fr_l1_1617$AvgHProbPN <- with(fr_l1_1617, round(1/BbAvH, 4))
fr_l1_1617$AvgDProbPN <- with(fr_l1_1617, round(1/BbAvD, 4))
fr_l1_1617$AvgAProbPN <- with(fr_l1_1617, round(1/BbAvA, 4))

fr_l1_1617$Overround <- with(fr_l1_1617, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
fr_l1_1617$AvgHProb <- with(fr_l1_1617, round(AvgHProbPN/Overround,4))
fr_l1_1617$AvgDProb <- with(fr_l1_1617, round(AvgDProbPN/Overround,4))
fr_l1_1617$AvgAProb <- with(fr_l1_1617, round(AvgAProbPN/Overround,4))
```

This provides us with columns of consensus probabilities: `AvgHProb`, `AvgDProb`, and `AvgAProb`.

IDA is begun by computing the mean consensus probabilities $\mu_m$ and respective STANDARD DEVIATION $\sigma_m$ for each market $m$. The results, along with the observed probability of each outcome, to four decimal places, are shown in Table 2.1.

Table 2.1: Basic calculations for the French Ligue Une, 2016/17 season, forming part of the initial data analysis.

| | *Home Win* | *Draw* | *Away Win* |
|---|---|---|---|
| Mean Consensus Probability $\mu_m$ | 0.4414 | 0.2701 | 0.2886 |
| Observed Probability | 0.4895 | 0.2474 | 0.2632 |
| Consensus Standard Deviation $\sigma_m$ | 0.1540 | 0.0469 | 0.1366 |

The table implies the mean consensus probabilities are similar to the observed probabilities: less than 0.05 off for each outcome. Interestingly, the standard deviation of a Draw is much lower than the standard deviation for both Home and Away Wins, indicating that the variation in the odds offered for Draws are much lower than those for a clear winner. To visualise these distributions, histograms of the consensus probabilities are produced and shown in Figure 2.1.

The figure shows that the consensus probabilities of a Home Win are symmetrically distributed around the mean $\mu_{\text{H}} = 0.44$, with a distribution similar to the NORMAL DISTRIBUTION bell curve. The consensus probabilities of an Away Win are positively- (or right-) SKEWED suggesting a greater proportion of measurements are greater than/lie to the right of the peak value (Mendenhall, Beaver, and Beaver, 2013). The data will, therefore, include a few unusually large measurements. Contextually, these could be the league leaders playing away against a poor performing side.

The graph of the consensus probabilities of a Draw assists the finding (based on $\sigma_{\text{D}}$) that the variation of consensus probabilities of a Draw is very low: Over 225 (of a total of $n = 380$ matches) of the games lie in the same bin[1] with no values at all recorded above 0.4.

[1]The width of all bins, or groups of data points, across all three histograms is 0.05.

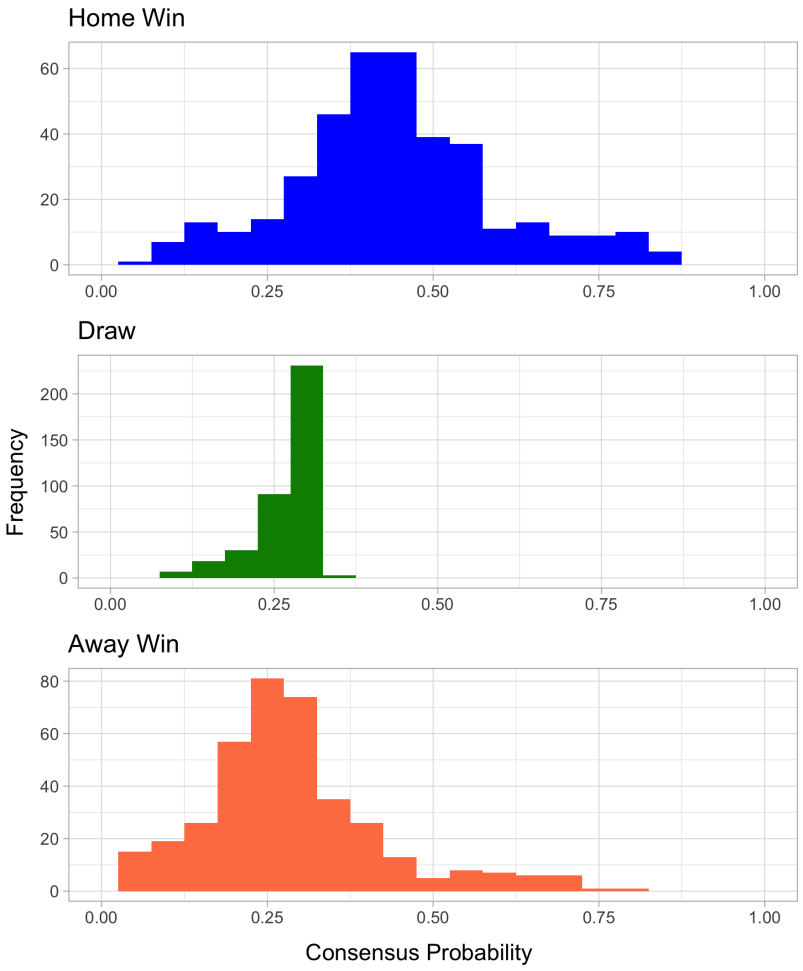


Figure 2.1: Histograms of the consensus probabilities for each outcome, in the French Ligue Une, 2016/17 season.

The rationale behind conducting IDA was to answer the four questions set out previously about the quality of the data and measurements; whether the implementation of the study fulfilled the intention of the research design; and the characteristics of the sample (Adèr, 2008). The quality of the data and measurements are both acceptable for this project: no measurements are missing, computations were completed, and plots produced without any issues. The question of the study will be able to be answered: access to the information required is available (consensus odds and actual results). The final question, about the characteristics of the sample, can be answered by saying that the data suggests the mean consensus probabilities are roughly equal to the observed probability for each outcome ($\mathbb{P}_{\text{cons}} \approx \mathbb{P}_{\text{obs}}$); bookmakers offer less variation on their odds for a Draw than for a clear winner; and the consensus probability of a Home Win is symmetric, whereas the consensus probability of an Away Win is positively-skewed.

## 2.2 Exploratory Data Analysis

In this section, the data is explored using various techniques motivated by four 'major themes'.

- **Displays**: reveal major features, outliers, non-linearities, discontinuities, skewness, ETC. that calculations such as means, standard deviations and least square regressions cannot show.
- **Residuals**: defined as the observed data minus the fitted data, a clear pattern in the plot of the residuals vs. fitted values indicates improvement is possible.
- **Resistance**: dealing with outliers. If they are found, parallel tests (one with outliers; one without) can be ran and compared, testing the resistance of the data to the outliers (not dissimilar to the idea of statistical LEVERAGE).
- **Transformations**: does adding a transformation, such as taking the $n$-th root, logarithms, logits/probits, ETC. allow sense to be made of the data? (Hoaglin, 1977)

The data must first be read in and cleaned (as six leagues and 15 seasons—90 datasets—are being considered, it is efficient to use a `for` loop to do this). The reasons for choosing these seasons are explained in Section 1.3. The code to do so—including finding the underlying probabilities, and to normalising to find consensus probabilities—is below.

```
countries <- c("de", "en", "es", "fr", "it", "po")
co.we <- c("D1", "E0", "SP1", "F1", "I1", "P1")
#n.b. The Premier League's code is 0; other countries are 1.
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112", "1213", "1314",
    "1415", "1516", "1617", "1718", "1819", "1920")
eliteTemp <- NULL; elite <- NULL
for (i in seasons){
   for (j in 1:6){
       eliteTemp <- read.csv(paste0('https://www.football-data.co.uk/mmz4281/', i, '/',
           co.we[j],'.csv'), fileEncoding = 'latin1')
       eliteTemp$Country <- with(eliteTemp, countries[j])
       eliteTemp$Season <- with(eliteTemp, i)
       if (i=="1920"){
           eliteTemp$BbAvH<-eliteTemp$AvgH; eliteTemp$BbAvA<-eliteTemp$AvgA
           eliteTemp$BbAvD<-eliteTemp$AvgD
       }
       else{}
       eliteTemp <- eliteTemp[ ,c("Div", "Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG",
           "FTR", "BbAvH", "BbAvD", "BbAvA", "Country", "Season")]
       elite <- rbind(elite, eliteTemp)
   }
}
elite <- na.omit(elite)

#Finding underlying probabilities:
#Pre-Normalised Probabilities
elite$AvgHProbPN <- with(elite, round(1/BbAvH, 4))
elite$AvgDProbPN <- with(elite, round(1/BbAvD, 4))
elite$AvgAProbPN <- with(elite, round(1/BbAvA, 4))
#To normalise them:
elite$overround<-with(elite, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
```

```
elite$AvgHProb <-with(elite, round(AvgHProbPN/overround, 4))
elite$AvgDProb <-with(elite, round(AvgDProbPN/overround, 4))
elite$AvgAProb <-with(elite, round(AvgAProbPN/overround, 4))
```

The number of matches in the dataset $N = 31,346$; as this is large, the CENTRAL LIMIT THEOREM (CLT) can be applied, allowing the assumption that the mean of the random variables (contextually, matches) follows the Normal distribution. For later analysis, the *correct* probability (and the natural logarithm of it) (the bookmaker consensus probability of the event that was observed) and the two *incorrect* probabilities need to be found also.

The first step after cleaning the data is to compute the consensus mean probabilities and compare them to the observed probabilities for each market $m$. These are shown in Table 2.2.

Table 2.2: Basic calculations for the entire *elite* dataset, forming part of the exploratory data analysis.

| | *Home Win* | *Draw* | *Away Win* |
|---|---|---|---|
| Mean Consensus Probability $\mu_m$ | 0.4472 | 0.2620 | 0.2908 |
| Observed Probability | 0.4589 | 0.2566 | 0.2845 |
| Consensus Standard Deviation $\sigma_m$ | 0.1714 | 0.0478 | 0.1536 |

As seen in Table 2.1, the mean consensus probabilities are extremely close to the observed probabilities: the magnitude difference is than 0.02 for all three outcomes. Similarly, the standard deviations indicate that the consensus probabilities offered for a Draw vary significantly less than for those with a clear winner.

Visual analyses are used; namely, box, density, and tile plots. First, a boxplot for each outcome is created, to assess whether there is a significant difference, or there exists a trend, in the observed outcome for the bookmaker probabilities of each event. These are shown in Figure 2.2.

The figure implies there is no significant difference: whilst it appears there is a correlation—for games that end in Home Wins, the consensus probability of a Home Win is higher, and vice versa for Away Wins—this plot shows no strong significance; however, it does strongly reiterate the aforementioned point about variation (or lack thereof) of the consensus probabilities of a Draw. The Draw plot has a large number of outliers; these are considered later.

Instead of creating a binned histogram, as in Figure 2.1, a density plot is used, allowing for an easier visualisation of the data, and easier comparisons between leagues. Using KERNEL DENSITY ESTIMATION, the data that would be used to create a histogram is 'smoothed out' using a number of equally spaced points at which the density (rather than the frequency, as in a histogram) is estimated. The plot for Home and Away Wins together is achieved using the code below; Draws are plotted separately due to a different $y$ axis scale being used. This results in Figure 2.3.

```
ggplot(elite, aes(x=AvgHProb, color="Home Win")) + geom_density() +
    geom_density(data=elite, mapping=aes(x=AvgAProb, color="Away Win"), show.legend=T) +
    coord_cartesian(xlim=c(0,1)) + labs(title="Home and Away Wins", caption="Elite
    Leagues, 2005-2020", x="Consensus Probability", y="Density") + theme_light() +
    scale_color_manual(name="Market", values=c("Home Win" = "blue", "Away Win" = "coral"))
```

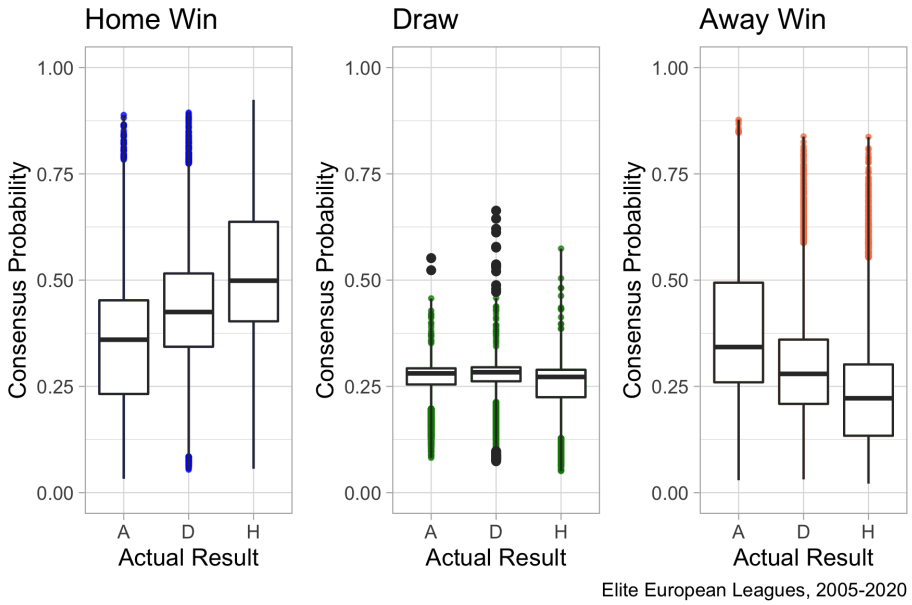


Figure 2.2: Boxplots of the consensus probabilities offered for each outcome.

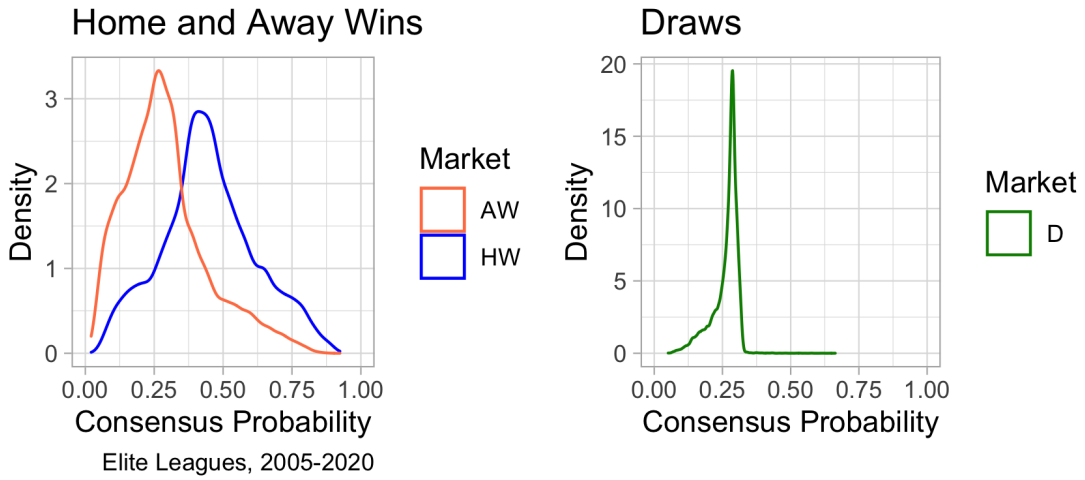


Figure 2.3: Density plots of the consensus probabilities offered for each outcome in the 1X2 market.

Interpreting this figure shows the trend shown from still holds: $\mathbb{P}_{\text{cons}}(\text{Home Win})$ is symmetrically distributed around the peak ($\mu_H = 0.45$); $\mathbb{P}_{\text{cons}}(\text{Away Win})$ has a positive skew; and $\mathbb{P}_{\text{cons}}(\text{Draw})$ has much less variation, with few matches above the peak.

An interpretation of this figure shows the trend from Figure 2.1 remains. The consensus proba-

bilities of a Home Win are symmetrically distributed around the mean ($\mu_H = 0.45$); the consensus probabilities of an Away Win have a positive skew; and the consensus probabilities of a Draw have much less variation, with few matches above the mean. Reviewing the latter, Table 2.3 shows matches with a consensus probability of a Draw greater than 0.6, where 0.6 is seven standard deviations away from the mean consensus probability of a Draw.

Table 2.3: Matches with consensus probability of a Draw, $\mathbb{P}_{\text{cons}}(\text{Draw}) > 0.6$

| *League* | *Date* | *Home Team* | *Away Team* | *Final Score* | $\mathbb{P}_{\text{cons}}$(Draw) |
|---|---|---|---|---|---|
| Serie A | 09/05/10 | Bologna | Catania | 1-1 | 0.6634 |
| Serie A | 08/05/11 | Bologna | Parma | 0-0 | 0.6445 |
| Serie A | 20/05/07 | Torino | Livorno | 0-0 | 0.6208 |
| Serie A | 03/04/11 | Chievo | Sampdoria | 0-0 | 0.6121 |

All these matches are both a) in the Italian Serie A; and b) in the late stages of the football season.[2] There are two possible reasons for this.

- The Italian Serie A has a history of match fixing in recent times: the CALCIOPOLI, which occurred during the 2004/05 and 2005/06 season involved Juventus, AC Milan, and Lazio, among others—three of Italy's largest clubs (Hafez, 2019); and in 2015, Catania's president was one of several arrested for match-fixing in Serie B matches (Gladwell, 2015).
- Due to these games occurring in the late stages, it is possible for a scenario to arise where both teams would benefit from a certain result.[3]

Whilst fixed matches would naturally impact the results, due to the small number of games impacted, it is unnecessary to exclude them from future analysis. As these games are all in the Serie A (a further investigation reveals, of the 81 matches in the dataset with a consensus probability of a Draw greater than 0.35, 75 (93%) were Italian matches), it makes sense to split this density plot into the different leagues, shown in Figure 2.4.

The Home Win (and to a lesser extent Away Win) plots imply leagues can be split into two categories: those with a *unimodal* distribution (one local MODE/peak), and those with a *trimodal* distribution (three local modes/peaks). The latter group (England, Portugal, Spain) have a peak between 0 and 0.25 (very low probability of a Home/Away win, depending on the plot) and a peak between 0.75 and 1 (very high probability of a Home/Away win), whereas the former (Germany, France, Italy) only has the central peak. One reason could be competitive balance, investigated further in Section 2.6.

The final visual aid used in this section is a tile plot. Similar to a heat map, this will allow for a three-dimensional representation of data in the two-dimensional plane. The match goals will be plotted on the $x$ (Home goals) and $y$ (Away goals) axes, which allows for the full-time result to be shown, (tiles on the diagonal $x = y$, are Draws; upper triangle are Away Wins $y > x$; lower triangle are Home Wins $x > y$) and the *magnitude* of the result, or how *convincing* the result is: a match further from the diagonal has a greater disparity in goals, and so can be considered a more convincing win. Representing the $z$ axis, each square will be shaded in with the mean *correct* consensus probability for that result: with a low consensus probability, the square will be lighter.

[2] The season normally starts in August and ends in May (The Football Association Premier League Limited, 2019).

[3] The DISGRACE OF GIJÓN in the 1982 World Cup is a particularly famous example of this (See Appendix A.2, Definition 5).

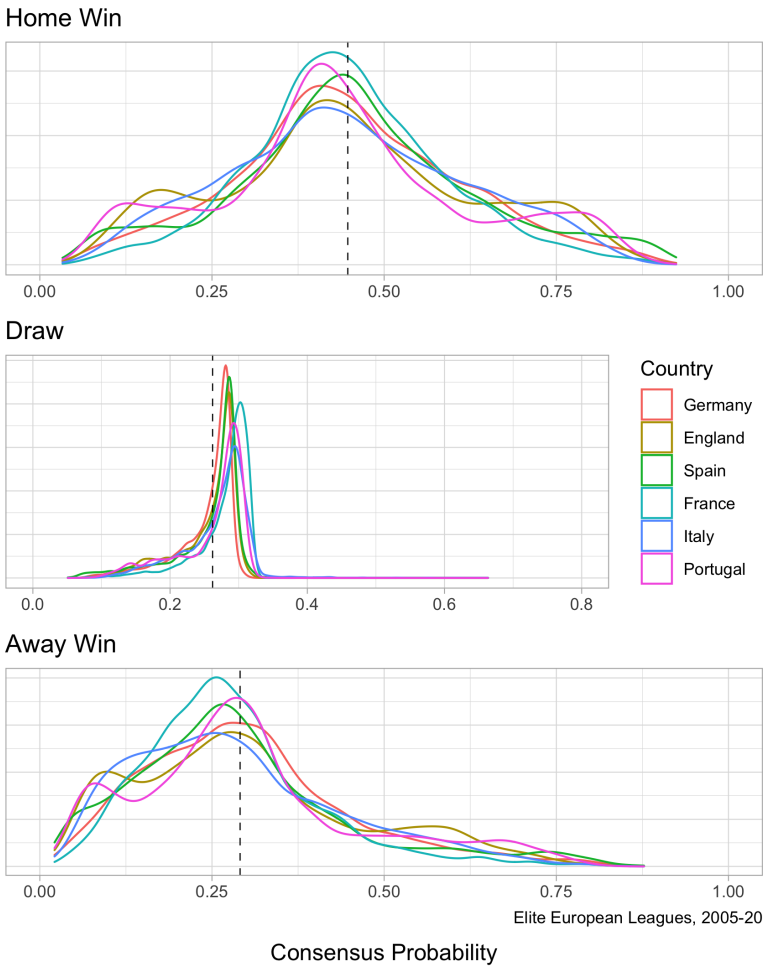


Figure 2.4: Density plots of the consensus probabilities offered for each outcome of the 1X2 market, split by league.

It is expected these to be closer to the diagonal. Due to a low number of extreme results where more than 5 goals are scored by a team, these are grouped these into a 5+ tile.[4] The bin sizes (number of matches with a given result) are given in Table 2.4.

The results on the diagonal all have a similar consensus probability, around 0.3. This is unsurprising, given the low variation in the consensus probabilities of Draws. Considering the lower triangle (Home Wins) first, the pattern expected holds: that is, tiles furthest from the diagonal are darker. Interestingly, the darkest tile is for a 5+ – 2 Home Win; the bin size for this is low, however, with $n = 106$ matches.[5] Whilst there are only 12 games in the 5+ – 4 tile, it is striking that it has one of the lowest correct consensus probabilities of all the tiles. Considering the upper triangle

[4]The highest scoring draws were 5–5, occurring twice: Lyon vs. Marseille (2009) and West Bromich Albion vs. Manchester United (2013).

[5]0.0034% of the total dataset.

Table 2.4: The bin size for each tile in Figure 2.5.

| Away Goals | Home Goals: 0 | 1 | 2 | 3 | 4 | 5+ |
|---|---|---|---|---|---|---|
| 5+ | 119 | 105 | 67 | 26 | 7 | 2 |
| 4 | 257 | 294 | 174 | 63 | 36 | 12 |
| 3 | 636 | 869 | 547 | 299 | 109 | 45 |
| 2 | 1451 | 2003 | 1508 | 712 | 265 | 106 |
| 1 | 2299 | 3688 | 2756 | 1375 | 528 | 281 |
| 0 | 2510 | 3342 | 2575 | 1371 | 594 | 315 |

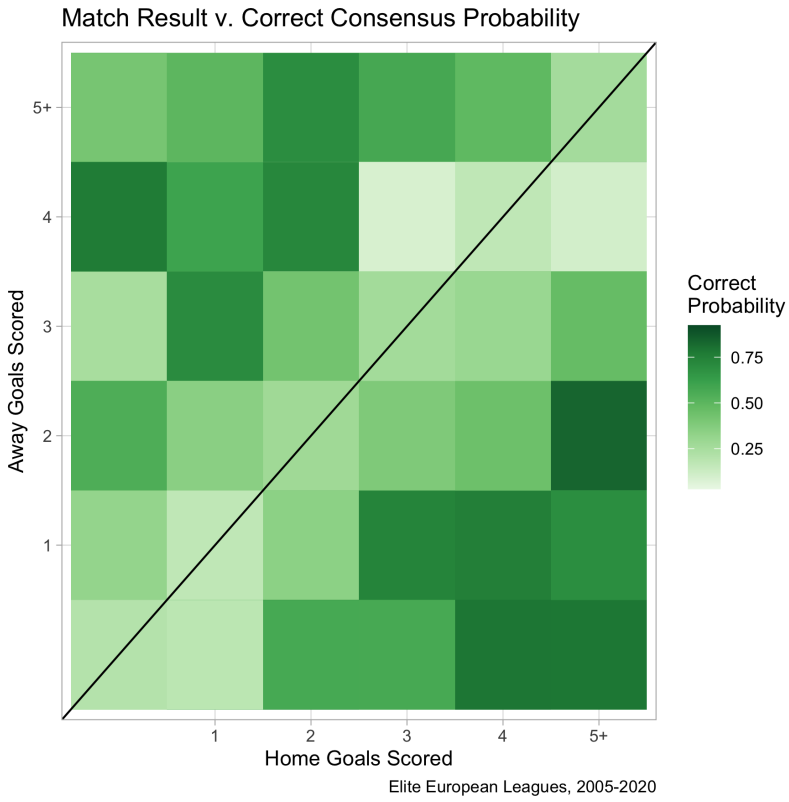


Figure 2.5: Tile plot of the correct consensus probability for each possible result.

(Away Wins), the pattern is not as consistent as with Home Wins, but there is still evidence of it, with the highest probabilities being further from the diagonal. As a whole, the figure is implying games with greater disparity in full-time goals (i.e., more convincing wins) have a higher bookmaker consensus probability of the correct result.

It is worth noting that more ideal measures of how convincing a match is exist, such as 'Expected Goals' (xG), a metric commonly used to evaluate the offensive-effectiveness of players, which can be used to create an 'Expected Results' and 'Expected League Tables', by giving each shot a theoretical

goal value based on historical shots from the same location, angle, position of defenders, goalkeeper positioning, ETC. (Rathke, 2017; footballxg.com, n.d.).

EDA was conducted to check four themes: displays, residuals, resistance, and transformations.

- Displays: Whilst there was no discontinuity–as in the IDA in Section 2.1—the consensus probabilities of Away Wins are positively/right-skewed, and of Draws have a very low VARIANCE. The outlying matches with a large $\mathbb{P}_{\text{cons}}(\text{Draw})$ were assessed.
- Residuals & Resistance: Models have not been created (this is done in Section 2.3), so the residuals cannot be plotted.
- Transformations: The data makes sense, and is as expected. There is nothing in the plots to suggest the need of, or to justify the use of, any transformations.

## 2.3 Correlation Analysis and Linear Model Creation

In this section, the coefficient of determination $R^2$ and the root-mean-square-error RMSE are found. This is done by creating a linear model to predict the observed probability (denoted $\mathfrak{O}$) from a given consensus probability ($\mathfrak{C}$). Linear models are chosen as the 'ideal' relationship between the observed and consensus probabilities is a $y = x$ line, where a bookmaker prediction of 0.25, for example, would correlate with an observed probability of 0.25. For the $i^{\text{th}}$ match (observation), the model created, in a set market, is of the form given in Equation 2.1, where $x_0$ is the intercept, $x_1$ is the *slope*, or gradient, of the line, $\mathfrak{C}_i$ is the consensus probability of the $i^{\text{th}}$ match, and $\epsilon_i$ is the error (residual) for that match.

$$\mathfrak{O}_i = x_0 + x_1 \cdot \mathfrak{C}_i + \epsilon_i \tag{2.1}$$

The RMSE for each game is defined as the square root of the mean residual point (the distance from each data point to the linear model created) squared. $R^2$ lies between 0 and 1, and is defined in Appendix A, Definition 4. High levels of correlation will result in a high $R^2$ and accurate models result in low RMSE values (Mendenhall, Beaver, and Beaver, 2013). There are alternatives to RMSE, such as mean square error, mean absolute error, median absolute error, and computing the distance from a point to the 'ideal' line: RMSE is more sensitive to outliers than mean and median absolute error; RMSE is chosen due to a relatively low amount of variation in values,[6] and its greater theoretical relevance (Hyndman and Koehler, 2006). $R^2$ describes the percentage of the variation in a dependant variable—in this case, the observed probability—due to a predictor—the bookmaker consensus probability (Draper and Smith, 1998).

The slope of the line can be used as an indicator of fit, too. As the 'ideal' model is known, a comparison between the fitted model $x_1$ and the ideal slope (1) can be made. The magnitude or percentage deviation could be used to check for discrepancy, regardless of whether the fitted model is steeper or shallower, which counters problems that may occur due to a pattern that oscillated above and below the ideal, however, the strict slope will be considered in this project, as such a trend would be of interest.

In addition to an overall $R^2$, RMSE, and slope, the three values for each *elite* league will be found, and used in Section 2.5.

[6] A caveat of RMSE is that it gives more weight to large errors, as the errors are squared. RMSE therefore is more useful when large errors are 'undesireable' (Wesner, 2016).

This is conducted by *binning* matches, where the data is partitioned into groups (bins), with a sufficient amount to 'capture the major features in the data while ignoring fine details' (Knuth, 2006). Games with similar consensus probabilities for a given outcome will be binned together, before the observed proportion of games in the bin with that outcome is found, which is referred to as the observed probability $\mathbb{P}_{\text{obs}}(\text{Outcome})$. This is used to find the consensus vs. observed probabilities in future analysis (plots and model creations). In R, bins are created using the `cut` function, and the `tapply` function is used to find the mean value in each bin.[7] The R code used for the Home Wins is below. 124 breaks (cut-points) are chosen, meaning that each bin has over 250 matches. Doing this for each outcome and overall results in the $R^2$, RMSE, and slope values presented in Table 2.5.

```
elite$AvgHProb.cut <- cut(elite$AvgHProb, 124, include.lowest=T)
#First, we cut the data into 'bins' choosing 124 breaks
levels(elite$AvgHProb.cut) <- tapply(elite$AvgHProb, elite$AvgHProb.cut, mean)
#Tapply finds the mean of the bin, rather than taking the midpoint
elite.observed.probabilites.TabH <- prop.table(table(elite$FTR, elite$AvgHProb.cut),
    2)[c(1, 2, 3),]
#The c(1,2,3) will remove any extra (blank) rows
elite.observed.probabilites.H <- elite.observed.probabilites.TabH[3,]
#[n,]; if n = : 1 Away; 2 Draw; 3 Home (alphabetic)
elite.bookmaker.probabilites.H <- as.numeric(names(elite.observed.probabilites.H))
```

Table 2.5: $R^2$, RMSE, and Slope values for the *elite* leagues, 2005–20.

| | *Home Win* | *Draw* | *Away Win* | *Overall* |
|---|---|---|---|---|
| $R^2$ | 0.98665 | 0.52008 | 0.96411 | 0.83832 |
| RMSE | 0.03241 | 0.20767 | 0.05166 | 0.11768 |
| Slope | 1.08278 | 1.38142 | 1.09182 | 1.11148 |

These values show that, for the Home and Away Win models, a strong correlation ($> 95\%$) exists between the consensus and observed probabilities.[8] In addition, the RMSE values for the two are very low: 0.03 for Home Wins and 0.05 for Away Wins, and the slopes are both less than 0.1 off the ideal. For Draws, $R^2 = 52\%$, RMSE $= 0.21$, and the slope is 1.38, indicating that the consensus probabilities, and thus the bookmaker's odds, are not as accurate than they are as for clear results; this explains why the overall values indicate worse performance than Home and Away Wins, with $R^2 = 83.8\%$, RMSE $= 0.12$, and the slope is 1.11.

A further demonstration of this point is given in Figure 2.6: a scatter plot of the observed probability vs. bookmaker consensus probability using the bins created. This is presented along with the linear models (and their corresponding 95% CONFIDENCE INTERVALS (CIs) in light grey). The 'ideal' $x = y$ line is included for reference (dashed black).

The figure affirms the observations from the values in Table 2.5, suggesting bookmaker accuracy for Home and Away Wins is extremely high. Both the 95% CIs are extremely small, with no major outliers. The plot reiterated the poor bookmaker performance for Draws, with a large 95% CI. It is also worth noting that the accuracy is higher for $\mathbb{P}_{\text{cons}}(\text{Draw}) \in (0, 0.3)$ (within the range 0 to 0.3) and remarkably poor for consensus probabilities above this range: this coincides with the straight

[7] The default is the midpoint (RDocumentation, n.d.).
[8] For a perfect correlation, $R^2 = 1$, 100% (Mendenhall, Beaver, and Beaver, 2013).

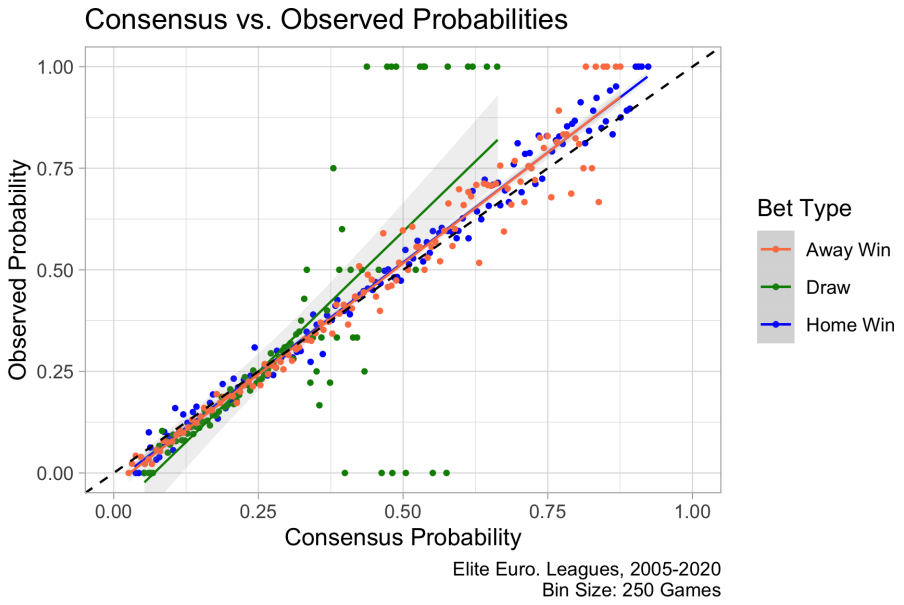


Figure 2.6: Scatter plot of the linear models created, showing the consensus probabilities vs. the observed probabilities.

drop in density in consensus probabilities of a Draw in Figure 2.3. The values on the $y = 1$ line—where all games in the bin were Draws—are likely to be due to the aforementioned match fixing or mutually-beneficial results (Section 2.2).

CIs are calculated by Equation 2.2 (Mendenhall, Beaver, and Beaver, 2013), where $\bar{x}$ is the point estimate, $\mu$ is the true value, and $Z_{\alpha=0.025} = 1.96$ is the $t-$statistic as $N$ approaches infinity (this can be assumed with a sufficiently large $N$). The large CI for the Draw model indicates the STANDARD ERROR (SE) is high; the SE values for each model are given in Table 2.6; the models created in this section, which can be used to predict the observed probability $\mathfrak{O}$ from a consensus probability $\mathfrak{C}$, are in in Equations 2.3 to 2.6.

$$\mathbb{P}\Big(\mu \in \big(\bar{x} \pm Z_{\alpha=0.025} \cdot SE\big)\Big) = 0.95 \tag{2.2}$$

$$\begin{aligned}
\mathfrak{O}_H &= -0.02323 + 1.08278 \cdot \mathfrak{C}_H && (2.3)\\
\mathfrak{O}_D &= -0.09646 + 1.38142 \cdot \mathfrak{C}_D && (2.4)\\
\mathfrak{O}_A &= -0.03018 + 1.09182 \cdot \mathfrak{C}_A && (2.5)\\
\mathfrak{O}_{\text{Overall}} &= -0.03194 + 1.11148 \cdot \mathfrak{C}_{\text{Overall}} && (2.6)
\end{aligned}$$

Table 2.6: The standard error SE of the linear model created for each outcome.

| | Model | | | |
|---|---|---|---|---|
| | *Home Wins* | *Draws* | *Away Wins* | *Overall* |
| SE | 0.011404 | 0.13835 | 0.019151 | 0.02651 |

## 2.4 Measuring Predictive Performance

A problem with $R^2$, RMSE, and the slope is that they are dependent on the bin size chosen to create the model. Two measures for calculating predictive performance are now considered that are independent of bin size. $P_1$ and $P_2$ are defined in Equations 2.7 and 2.8 respectively (Owen, 2009), with $N$ being the number of matches in the sample; $k$ being the match number; $\mathbb{P}(O_k)$ being the correct/observed bookmaker probability for match $k$; and $\mathbb{P}(I_{1k})$, $\mathbb{P}(I_{2k})$ being the two incorrect/not observed bookmaker probabilities for match $k$: for example, if match $m$ finished in a win for the home side, $\mathbb{P}(O_m) = \mathbb{P}_{\text{cons}}(\text{Home Win})$, and $\mathbb{P}(I_{1m})$ and $\mathbb{P}(I_{2m})$ are the consensus probabilities of an Away Win and of a Draw (it does not matter which is considered incorrect outcome 1 or 2).

$$P_1 = \exp\left\{\frac{1}{N}\sum_{k=1}^{N}\log_e\left[\mathbb{P}(O_k)\right]\right\}, \quad k \in [1,\ N] \tag{2.7}$$

$$P_2 = \frac{1}{N}\sum_{k=1}^{N}\left\{\left[1-\mathbb{P}(O_k)\right]^2 + \mathbb{P}(I_{1k})^2 + \mathbb{P}(I_{2k})^2\right\}, \quad k \in [1,\ N] \tag{2.8}$$

Better predictive performance is given by higher $P_1$ and lower $P_2$ values, though are not well-suited for stand-alone interpretation. Instead, these values will be used as a comparative measure across leagues and seasons. The values for the entire *elite* dataset are given in Table 2.7.

Table 2.7: $P_1$ and $P_2$ values for the entire *elite* dataset.

| *Measure* | *Value* |
|---|---|
| $P_1$ | 0.3788607 |
| $P_2$ | 0.5776072 |

## 2.5 Comparing Leagues

The six *elite* leagues are compared in this section. To compare them, a near-identical analysis as in Sections 2.3 and 2.4 is carried out, to find each league's $R^2$, RMSE, $P_1$, and $P_2$ value. In the correlation analysis, a weight is added to the number of bins: the more varied Home and Away Wins are split into 20 bins (around 200 games per bin) and the less varied Draws are split into five (800). In R, a `for` loop is used to create a temporary linear model, where the $R^2$, RMSE, and slope values are extracted and added to an array where all league's values are stored. The results (and their respective ranking, $R^2$ and $P_1$ descending; RMSE and $P_2$ ascending, with higher ranking indicating better performance) are presented in Table 2.8. (N.B. the slope is not included in the ranking, or in future measures of 'accuracy' as it would give more weight to the linear model

than the $P$-values. The slope is omitted rather than the $R^2$ and RMSE because it does not assess correlation: a model could have a slope of 1 with poor results that would have a low $R^2$/high RMSE, or a perfectly-correlated line could have a shallow/steep slope. Therefore, confidence intervals were added to the plots in Section 2.3. A well-correlated slope shallower or steeper than the ideal could be due to bookmaker overround, investigated in Section 3.5., or due to bookmakers purposefully setting mispriced odds for profitability reasons: this would normally result in a shallower slope—as seen with the AH markets—with the high consensus probabilities (lower odds) occurring less than expected. The slope gives a good trend of such impacts.)

Table 2.8: $R^2$, RMSE, $P_1$, $P_2$, and Slope for all *elite* leagues.

| | $R^2$ | RMSE | $P_1$ | $P_2$ | Slope | *Average Rank* |
|---|---|---|---|---|---|---|
| *Germany* | 0.95489<br>6th | 0.05472<br>6th | 0.36958<br>5th | 0.59429<br>5th | 1.01470<br>— | 5.50 |
| *England* | 0.96773<br>5th | 0.04783<br>5th | 0.38454<br>2nd | 0.56661<br>2nd | 1.07368<br>— | 3.50 |
| *Spain* | 0.96950<br>4th | 0.04711<br>4th | 0.38267<br>4th | 0.57029<br>3rd | 1.06122<br>— | 3.75 |
| *France* | 0.97938<br>3rd | 0.03845<br>3rd | 0.36591<br>6th | 0.60248<br>6th | 1.09520<br>— | 4.50 |
| *Italy* | 0.98599<br>1st | 0.03297<br>1st | 0.38287<br>3rd | 0.57085<br>4th | 1.18288<br>— | 2.25 |
| *Portugal* | 0.98216<br>2nd | 0.03642<br>2nd | 0.38894<br>1st | 0.55963<br>1st | 1.14625<br>— | 1.50 |

Looking at the average ranking, the bookmakers performed most accurately in the Portuguese and Italian leagues, followed by the English and Spanish leagues, and least accurately in the French and German leagues. An investigation into whether this is due to competitive balance is carried out in Section 2.6.

## 2.6 The Effect of Competitive Balance on Bookmaker Accuracy

### 2.6.1 Defining Competitive Balance

Competitive balance is a concept that weighs heavily on economics. It is defined, by the Cambridge Dictionary (n.d.), as 'the situation in which no one business of a group of competing businesses has an unfair advantage over the others,' with a monopoly being a situation in which competitive balance does not exist.

There is a weight of research into competitive balance in American sports, with all of the 'big four' leagues[9] and the MLS (Major League Soccer: the top-tier football league in the USA) having some form of salary cap to promote balance (Slowinski, 2012; NBA.com, 2020; Rosen, 2020; Barrabi, 2020; Goal, 2019). The nature of competitive balance in baseball is that 'competitors must be of approximately equal 'size' if any are to be successful; this seems to be a unique attribute of professional competitive sports' (Rottenberg, 1956).

In the English Premier League 2019/20 season, Sheffield United (with an average player salary of $0.91 million: the lowest in the league) finished in ninth position, only one place behind Arsenal (with an average player salary of $5.99 million, over six times higher than Sheffield United's, and the fourth highest in the league). The two matches between the sides ended in a 1-1 draw (18$^{\text{th}}$ January, 2020) and a 1-0 win for Sheffield United (21$^{\text{st}}$ October, 2019) (Lange, 2021; Limited, 2020). Similarly, the 2015/16 season, the league winners—Leicester City—had a turnover of £129 million, and paid £80 million in wages, whilst Chelsea—with a turnover of £335 million, and paid £224 million in wages—finished tenth (Conn, 2017).

This is a pattern replicated across Europe. In 2020, Statista published a ranking of the German Bundesliga teams by market (transfer) value prior to the 2020/21 season. FC Bayern Munich had a value of €875m whilst DSC Arminia Bielefeld had a value of €47m (Lange, 2020); the clubs drew 3-3 at Munich's stadium.

This shows that European football may not, in terms of competitive balance, follow the same guides that American sport does, and thus research into American competitive balance cannot be applied to this project.

### 2.6.2 Quantifying Competitive Balance

Three measures of quantifying competitive balance will be used. The National Measure of Seasonal Imbalance (NAMSI), the Top $K$ Ranking ($\kappa$), and the Gini coefficient.

The NAMSI, shown in equation 2.9, is (for a league of $n$ teams), the ratio between two standard deviations: $\sigma_{\text{Season}}$ the observed standard deviation of each team's winning percentage (for team $i$, $W_i$); and $\sigma_{\text{Certainty}}$ the theoretical standard deviation of a certain season, where the team in first place wins every game, the team in second wins every game except those against the team in first, ETC., and the team in $i^{\text{th}}$ place loses to all teams that finish above them in the table, and defeat all those below (for team $i$, $C_i$). The NAMSI statistic lies in the range between 0 and 1, with a lower value corresponding to lower seasonal imbalance (higher balance) (Goossens, 2005).

$$\text{NAMSI} = \frac{\sigma_{\text{Season}}}{\sigma_{\text{Certainty}}} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(W_i - 0.5)^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(C_i - 0.5)^2}{n}}} \tag{2.9}$$

[9]The MLB (baseball), NBA (basketball), NFL (American football), NHL (ice hockey).

NAMSI is a 'static measure since it only looks at one season independently of other seasons.' In football leagues, a poorly competitively balanced league system would have the same few teams competing for the same places each season. A fluid measure is introduced to combat this: the Top $K$ ranking, denoted $\kappa$ (kappa). K is chosen to be equal to three.[10] This is the number of unique teams entering the top three in three consecutive years, with data from 1963/64 to 2004/05, with a value of 3 showing perfect imbalance, and 9 showing perfect balance. (Goossens, 2005).

A further method of quantifying imbalance is the Gini coefficient, $G$, which is often used to compare wealth inequality between different nations (The World Bank, 2018). It measures the ratio of the area between the Lorenz Curve of the country and the $y = x$ (perfect equality) line.[11].

In Table 2.9, the NAMSI, $\kappa$ and Gini coefficients are shown (Goossens, 2005) with the *elite* leagues ranked (NAMSI, Gini descending; $\kappa$ ascending: higher ranking indicates higher imbalance in the league), and other top-tier European leagues included for reference.

Table 2.9: Competitive balance statistics for a range of European leagues, across the 1963/64–2004/05 Seasons (Goossens, 2005).

| | NAMSI | | $\kappa$ | | Gini | | *Average Rank* |
|---|---|---|---|---|---|---|---|
| *Germany* | 0.374 | 3rd | 5.71 | 4th | 0.723 | 6th | 4.3 |
| *England* | 0.372 | 4th | 5.79 | 5th | 0.826 | 3rd | 4 |
| *Spain* | 0.364 | 5th | 5.07 | 2nd | 0.861 | 2nd | 3 |
| *France* | 0.342 | 6th | 6.00 | 6th | 0.784 | 4th | 5.3 |
| *Italy* | 0.418 | 2nd | 5.36 | 3rd | 0.737 | 5th | 3.3 |
| *Portugal* | 0.505 | 1st | 4.07 | 1st | 0.898 | 1st | 1 |
| *Belgium* | 0.452 | — | 5.07 | — | 0.801 | — | — |
| *Denmark* | 0.412 | — | 6.43 | — | 0.581 | — | — |
| *Greece* | 0.488 | — | 4.14 | — | 0.870 | — | — |
| *The Netherlands* | 0.494 | — | 4.36 | — | 0.888 | — | — |
| *Sweden* | 0.410 | — | 6.07 | — | 0.692 | — | — |

Table 2.9 implies the Portuguese is the least competitively balanced by all measures; the Italian league has a high NAMSI and low $\kappa$ value, indicating low competitive balance, but—contrastingly—a low Gini coefficient. By the NAMSI and $\kappa$ values, the French Ligue Une was the most competitive league, and by the Gini coefficient, the German Bundesliga was.

These values are reinforced by Ramchandani (2012), who found—using six measures[12]—that in the 2010/11 season only, the French league was the most competitive, followed by the English, German, Spanish, Italian, and finally, the Portuguese (the least competitive) leagues.

Comparing these values to Table 2.8, the leagues with the greatest bookmaker performance (Portugal and Italy) have the worst competitive balance, and vice versa: Germany and France have

[10]This value, whilst arbitrary, is chosen 'because in most European countries it are two or three teams that in general are considered to be dominant. Taking up more teams underrates the dominance since the top 4 and 5 often change' (Goossens, 2005).

[11]The Lorenz Curve is a graphical measure showing the overall income distribution: on the $x$ axis is the cumulated percent of the population, from poorest to richest; on the $y$ axis is the percent of the total wealth of the country held by this $x$% (Lorenz, 1905). For the use in football, income/total wealth is replaced with winning percentages.

[12]Inter-quartile range; top-bottom quartile gap; coefficient of variance; top 25% concentration ratio; top 50% concentration ratio; and the Hirfindahl-Hirschman index.

the lowest bookmaker performance, and the highest competitive balance.

If Figure 2.4 is considered—which implies the Spanish, Portuguese and English leagues follow a trimodal distribution; the French, German and Italian leagues follow a unimodal distribution—a link between competitive balance and the distribution can be inferred. Portugal's imbalanced Primiera Liga has a trimodal distribution and higher levels of bookmaker accuracy; both Spain's La Liga and England's Premier League also follow the trimodal distribution, and had similar levels of balance and accuracy. The well-balanced German Bundesliga and French Ligue Une both have a unimodal distribution and lower levels of bookmaker accuracy. The only exception is the Italian Serie A, with low balance and high accuracy but a unimodal distribution.

Boscá et al. (2009) conducted a study into the Italian and Spanish football, analysing styles of play (offensive or defensive), and showed that whilst the Italian league requires strong defensive efficiency to achieve a high ranking; the contrary is true in Spain, where the 'best-rewarded strategy consists in improving offensive efficiency.' A relationship between the style of play and the bookmaker accuracy could be investigated further.

## 2.7 Comparing Seasons

In addition to running the by-league analysis, a by-season analysis is conducted to investigate whether any changes have taken place over time. The values are presented in Table 2.10 and plotted in Figure 2.7.

Table 2.10: $R^2$, RMSE, $P_1$, $P_2$, and Slope for the *elite* league data, split by season.

| *Season* | $R^2$ | RMSE | $P_1$ | $P_2$ | Slope |
|---|---|---|---|---|---|
| 05/06 | 0.84576 | 0.10689 | 0.37340 | 0.58823 | 1.25361 |
| 06/07 | 0.80332 | 0.12249 | 0.36796 | 0.59849 | 1.24736 |
| 07/08 | 0.82067 | 0.10773 | 0.37081 | 0.59307 | 1.12600 |
| 08/09 | 0.89556 | 0.08401 | 0.37495 | 0.58519 | 1.26211 |
| 09/10 | 0.94893 | 0.06100 | 0.38075 | 0.57518 | 1.18172 |
| 10/11 | 0.89620 | 0.08000 | 0.36912 | 0.59602 | 1.07412 |
| 11/12 | 0.96151 | 0.04783 | 0.37626 | 0.58291 | 1.03563 |
| 12/13 | 0.90565 | 0.08169 | 0.37919 | 0.57617 | 1.15301 |
| 13/14 | 0.97634 | 0.03937 | 0.38644 | 0.56293 | 1.08188 |
| 14/15 | 0.81375 | 0.12699 | 0.37990 | 0.57565 | 1.09294 |
| 15/16 | 0.98860 | 0.02814 | 0.37731 | 0.58059 | 1.06431 |
| 16/17 | 0.97868 | 0.04035 | 0.39273 | 0.55283 | 1.11732 |
| 17/18 | 0.99342 | 0.02136 | 0.39202 | 0.55363 | 1.09199 |
| 18/19 | 0.78709 | 0.13288 | 0.38508 | 0.56514 | 1.06774 |
| 19/20 | 0.98955 | 0.02433 | 0.37715 | 0.57974 | 1.01129 |

The figure clearly shows that accuracy is improving over time, for each measure: $R^2$ and $P_1$ are increasing; RMSE and $P_2$ are decreasing; and the Slope is approaching the ideal $y = 1$ line. This is likely due to advances in the betting models used by bookmakers and increased information available, such as historical data and more metrics, E.G., Expected Goals (xG) (Rathke, 2017).

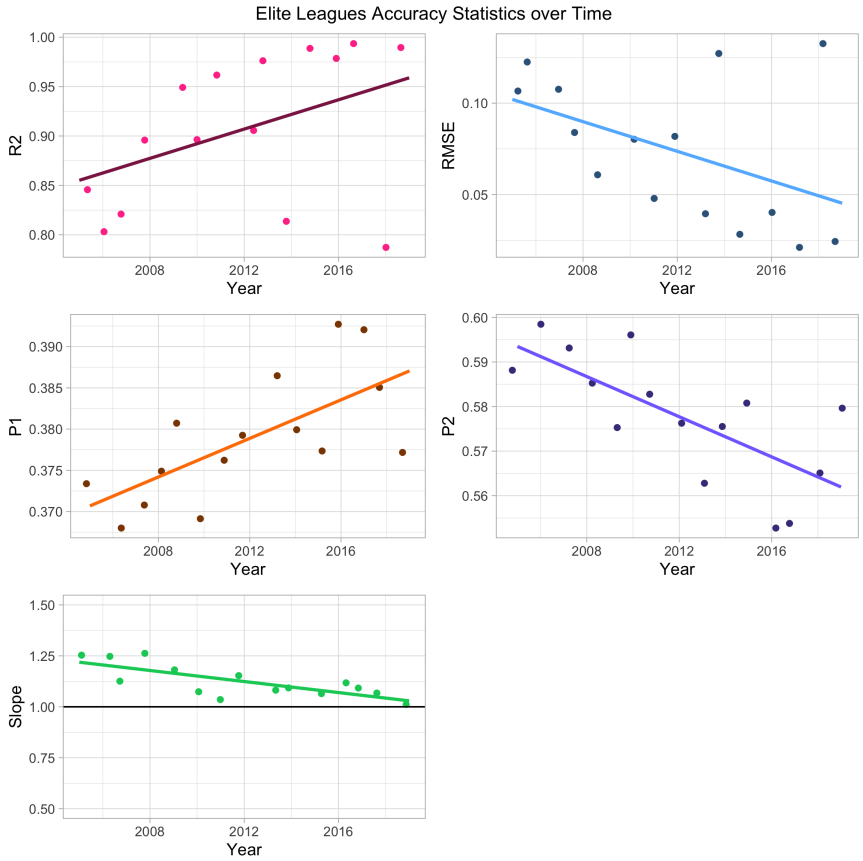


Figure 2.7: The variation of $R^2$, RMSE, $P_1$, $P_2$, and Slope over time in the *elite* leagues data.

## 2.8 Principal Component Analysis

The final analysis in this chapter is a pair of principal component analyses: firstly, by-league; secondly, by-season, used to group similar observations. Principal components analysis (PCA) is a tool used in multivariate analysis to simplify the number of variables a data set has (Wold, Esbensen, and Geladi, 1987), aiming to find 'lines and planes of closest fit to systems of points in space' (Pearson, 1901). Graphically PCA can be thought of as setting a 'new' pair of axes to a scatter plot: the 'new' $x$ axis being on the LINE OF BEST FIT and the 'new' $y$ axis being on the line of worst fit.

### 2.8.1 Choosing Which Components to Keep

The importance of components output will show how much variation is accounted for by each component, both individually and cumulatively from the 'most influential' to the 'least.' In this project, two methods for choosing which components to keep are used. N.B., it can be shown the eigenvalues are equal to the variances (Alto, 2019). First, the *Kaiser criterion*, which states to only keep a component if its variance (eigenvalue) is greater than 1 (Kaiser, 1974). The choice of 1 is

arbitrary and not recommended as a strict rule (Fabrigar et al., 1999), thus, a second criterion: the *Jolliffe criterion* is also applied: components with eigenvalues greater than 0.7 times the average eigenvalue are retained (Jolliffe, 1972). In both screeplots in this section, the Kaiser criterion is added.

### 2.8.2 By-League Principal Component Analysis

To conduct the first analysis—PCA by-league—three new variables are introduced: *Imbalance*, *Level of Attack*, and *Predictive Accuracy*. This is due to the 'small-$n$-large-$p$' problem,[13] where one has more predictors (in this case, there are eight: $R^2$, RMSE, $P_1$, $P_2$, NAMSI, $\kappa$, Gini, and the Level of Attack) than observations/samples (six leagues) (Ma and Dai, 2011). Thus, the number of predictors needs to be reduced to achieve meaningful results, so scaled averages of predictor groups are taken.

The Imbalance (`imbalance` in the code), found for each country $c$ using Equation 2.10, is a scaled average of the NAMSI $N$, inverse-$\kappa$ denoted $\tau$ (the inverse is taken so high scores in all indicate higher imbalance), and Gini coefficient $G$ (these variables were explained in Section 2.6) (Goossens, 2005).

$$\text{Imbalance}_c = \frac{1}{3}\left[\left(\frac{c_N - \mu_N}{\sigma_N}\right) + \left(\frac{c_\tau - \mu_\tau}{\sigma_\tau}\right) + \left(\frac{c_G - \mu_G}{\sigma_G}\right)\right] \tag{2.10}$$

The Level of Attack LA (`attack.league` in the code) for each country $c$ is found by Equation 2.11. This is computed in R by re-reading the dataset from `football-data.co.uk`, as done in Section 2.2, instead using the `FTHG`, `FTAG`, `HS` (Home Shots), and `AS` (Away Shots) columns. For $n$ seasons of data, the Level of Attack is defined as the average ratio of shots per game (HS+AS) to goals per game (HG+AG). (N.B., `football-data.co.uk` only has the shots data for the Portuguese Premiera Liga from the 2017/18 season onwards; it is thus assumed the Level of Attack in the league has been constant since 2005 at the level from 2017.)

$$\text{LA}_c = \frac{1}{n}\sum_{s=1}^{n}\left(\frac{\mu_{(HS+AS),\ s}}{\mu_{(HG+AG),\ s}}\right) \tag{2.11}$$

Similar to imbalance, the Predictive Accuracy PA (`predacc` in the code) is a scaled average of the predictive variables: $R^2$, RMSE, $P_1$ and $P_2$ (the inverse of RMSE and $P_2$ is taken, denoted $\zeta$ and $\theta$ respectively), found by Equation 2.12.

$$\text{PA}_c = \frac{1}{4}\left[\left(\frac{c_{R^2} - \mu_{R^2}}{\sigma_{R^2}}\right) + \left(\frac{c_\zeta - \mu_\zeta}{\sigma_\zeta}\right) + \left(\frac{c_{P_1} - \mu_{P_1}}{\sigma_{P_1}}\right) + \left(\frac{c_\theta - \mu_\theta}{\sigma_\theta}\right)\right] \tag{2.12}$$

Once these values are computed, PCA can be ran using the code below. This results in three principal components. The output from the code is in Table 2.11, where the *rotations* are the component values.

```
pc.league <- matrix(c(imbalance, attack.league, predacc), ncol=3, byrow=F)
colnames(pc.league) <- c("imbalance", "attack", "predacc")
rownames(pc.league) <- countries

league.model <- prcomp(pc.league)
league.model$rotation; summary(league.model)
```

[13] Also referred to as '$n << p$', 'small-$N$-large-$d$' in literature.

Table 2.11: By-League PCA values.

| **Importance of components:** | | | |
|---|---|---|---|
| Standard deviation | 1.166 | 0.5301 | 0.13284 |
| Proportion of Variance | 0.820 | 0.1694 | 0.01064 |
| Cumulative Proportion | 0.820 | 0.9894 | 1.00000 |
| **Component rotations:** | | | |
| | PC1 | PC2 | PC3 |
| `imbalance` | 0.7382378 | 0.4105132 | -0.5352419 |
| `attack` | -0.3206651 | -0.4845176 | -0.8138898 |
| `predacc` | 0.5934466 | -0.7724776 | 0.2260519 |

### Component Retention

By the both the Kaiser and Jolliffe criterions, only PC1 is retained. Figure 2.9 has the Kaiser criterion overlaid; the mean variance times 0.7 is found by:

```
> mean(c(1.166**2, 0.5301**2, 0.13284**2))*0.7
[1] 0.3869153
> 0.5301**2
[1] 0.281006
```

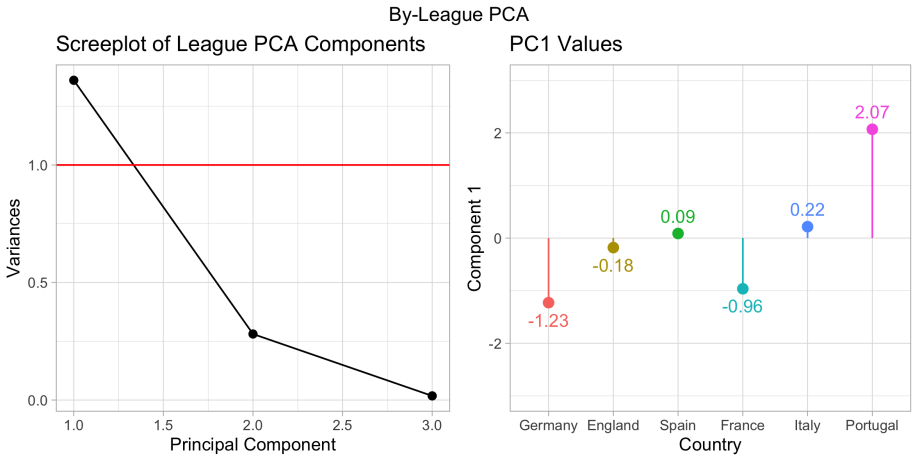


Figure 2.8: By-League PCA figures.

### Interpretation

PC1 can be considered a contrast between the Imbalance and Predictive Accuracy (PA), against the Level of Attack (LA) in a league (high values could be due to the league being defensive, or high imbalance and accuracy). A plot of the PC1 values is provided alongside the screeplot in Figure 2.8.

There are three groupings that can be made from this plot: the first, is the Portuguese league on its own, with the highest score (2.07) by far. Second are the three leagues near a score of 0, Italy (0.22), Spain (0.09), and England (-0.18). Finally, the most-negative group of leagues are the French (-0.96) and German (-1.23) leagues. This analysis appears has been dominated by the Imbalance and PA scores: literature has shown the Spanish (attacking) and Italian (defensive) leagues have opposing styles of play (Boscá et al., 2009), yet have a similar PC1 value. Both leagues show similar competitive balance statistics (Table 2.9) and accuracy statistics (Table 2.8) indicating the style of play hasn't affected the PC1 score enough to provide conclusive findings. Further analysis with more methods of analysing the style of play may provide better results.

### 2.8.3 By-Season Principal Component Analysis

Unlike with Section 2.8.2, there are more observations (15 seasons) than predictors (five: the four accuracy measures and the Level of Attack; competitive balance values are unavailable by-season), so there is no need to take scaled averages to reduce the number of predictors. Thus, PCA is conducted on the following variables: $R^2$, inverse RMSE ($\zeta$), $P_1$, inverse $P_2$ ($\theta$), and the Level of Attack LA for that season, found by Equation 2.13, where $c$ represents the country; $s$ the season.

$$\text{LA}_s = \frac{1}{n}\sum_{c=1}^{n}\left(\frac{\mu_{(HS+AS),\ c}}{\mu_{(HG+AG),\ c}}\right) \tag{2.13}$$

The code, including the creation (and scaling) of a dataframe, is below; the outputs are in Table 2.12, as well as a screeplot and a scatter plot of PC1 vs. PC2 in Figure 2.9.

```
pc.season <- matrix(c(rsqu.season, (1/rmse.season), p1.season, (1/p2.season),
    attack.season), ncol = 5, byrow=F)
colnames(pc.season) <- c("rsqu", "inv rmse", "p1", "inv p2", "attack")
rownames(pc.season) <- seasons
pc.season.sc <- scale(pc.season)

season.model <- prcomp(pc.season.sc)
summary(season.model); round(season.model$rotation,3)
```

Table 2.12: By-Season PCA values.

**Importance of components:**

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Standard deviation | 1.8112 | 0.9998 | 0.7909 | 0.30632 | 0.02448 |
| Proportion of Variance | 0.6561 | 0.1999 | 0.1251 | 0.01877 | 0.00012 |
| Cumulative Proportion | 0.6561 | 0.8560 | 0.9811 | 0.99988 | 1.00000 |

**Component rotations:**

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| $R^2$ | 0.409 | -0.624 | 0.199 | 0.636 | 0.013 |
| Inverse RMSE, $\zeta$ | 0.453 | -0.495 | -0.241 | -0.701 | -0.010 |
| $P_1$ | 0.489 | 0.364 | 0.360 | -0.055 | -0.704 |
| Inverse $P_2$, $\theta$ | 0.491 | 0.372 | 0.333 | -0.070 | 0.710 |
| Level of Attack | -0.383 | -0.308 | 0.813 | -0.310 | 0.014 |

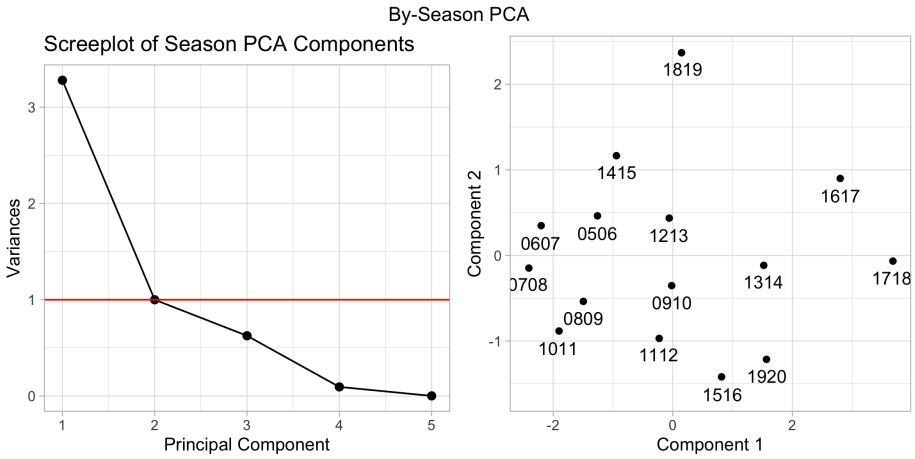


Figure 2.9: By-Season PCA figures.

### Component Retention

The importance of components output (Table 2.12) shows the first two components account for 85.6% of the cumulative variance. The Kaiser criterion on the screeplot (Figure 2.9) suggests to retain PC1, and that PC2 is borderline (on the criterion line). Considering the Jolliffe criterion, it is recommended to keep both PC1 and PC2:

```
> mean(c(1.8112**2, 0.9998**2, 0.7909**2, 0.30632**2, 0.002448**2)) * 0.7
[1] 0.6999169
> 0.9998**2 #Component 2
[1] 0.9996
> 0.7909**2 #Component 3
[1] 0.6255228
```

### Interpretation

PC1 can be interpreted as a contrast between the predictive variables ($R^2$, $\zeta$, $P_1$, $\theta$), and the level of attack, with high values being awarded to seasons with high levels of accuracy, and/or lower levels of attacking football.

PC2 can be interpreted as a contrast between $P_1$ and the inverse of $P_2$, and—mainly—$R^2$. Whilst both are statistics of accuracy, the two are not mutually agreeable: in Table 2.10, the 2018/19 season has the extreme poor-performing $R^2$ and RMSE, but performs on-trend for $P_1$ and $P_2$ (Figure 2.7).

In the PC1 vs. PC2 plot in Figure 2.9, the earlier years (2005/06 until 2012/13) are contained within the bottom-left (low PC1, low PC2) of the graph (bounded by $x = 0,\ y = 0.5$); in the bottom right are the 15/16, 19/20, 13/14 and 17/18 seasons, with low PC2, and high PC1: indicating lower attacking football (high PC1), and high levels of accuracy, especially the $R^2$ and RMSE measures, indicating both a trend of accuracy increasing over time (as mentioned in Section 2.7), and of more recent football seasons being more attacking. This is an area for future research, perhaps using a

larger dataset and more advanced measures of attacking styles of play.

## 2.9 Conclusion

In this chapter, it has been shown that the levels of bookmaker accuracy are high in the 1X2 Home Win and Away Win markets: in each of the *elite* European leagues, the coefficient of determination $R^2$ was above 95% with the RMSE below 0.055. From the scatter plot in Figure 2.6, the linear models for Home and Away wins have a much lower standard error (and therefore a narrower CI) than the model for Draws. This, along with the $R^2$ and RMSE for each result in Table 2.5, indicates that whilst bookmakers enjoy high accuracy with clear results, their predictions for Draws are poor, and have large room for improvement.

In addition, it has been shown that the accuracy is impacted by the competitive balance in each league: bookmakers perform better in countries in imbalanced leagues, such as Portugal's Primiera Liga and Italy's Serie A, than in balanced leagues, such as Germany's Bundesliga and France's Ligue Une.

---

In Chapter 3, the accuracy of bookmaker's odds in the English & Scottish leagues is investigated across multiple levels (rather than just the *elite* leagues) and betting markets (rather than just the 1X2 market).

# Chapter 3

# Assessing the accuracy of betting odds in the English & Scottish football league pyramids, from 2005 to 2020.

In this chapter, the accuracy of betting odds across multiple levels of football leagues are assessed. This is done by using the English & Scottish football league pyramids, with the Under/Over 2.5 Goals (UO) and Asian Handicap (AH) markets considered, in addition to the 1X2 market as in Chapter 2. Reasons for choosing these leagues are outlined in Section 1.1. Many of the same techniques, including conducting exploratory data analysis (EDA), correlation analysis, and assessing the predictive power will be utilised. The overround across the markets and leagues, a measure of the bookmaker commission, will also be assessed.

## 3.1 Exploratory Data Analysis

### Reading and cleaning the data

Initial data analysis, checking the data source, has already been done (Section 2.1), and thus exploratory data analysis can be conducted immediately. Again, a `for` loop is used to read the data directly from `football-data.co.uk`.

Once the data is imported, the `level` of the league is defined. Level 1 is defined as the English Premier League (EPL) and Championship, and the Scottish Premier League (SPL). The EPL and the SPL are the top-tier leagues in their pyramids; the Championship's average revenue per club is €33 million per club, per year (€13 million on average per year more than the SPL) (Ajadi et al., 2020) and so is included in this level. Level 2 is defined as the remaining fully professional leagues in the pyramids: these are the English Leagues One and Two, and the Scottish Championship. Finally, Level 3 is defined as the leagues with semi-professional sides in: the English Conference and Scottish Leagues One and Two. The Under/Over 2.5 Goals (UO) and Asian Handicap (AH) consensus odds are also read in.

After reading in the data, and adding the `level` using a `for` loop, the underlying and consensus probabilities (Section 1.1.2) are found using the same method as in Chapter 2. There are a couple of major mistakes in `football-data.co.uk`'s files with handicaps recorded, fixed below:

```
#Rangers had a -275 goal handicap vs. East Fife; assume this meant -2.75:
ensco$HomeHandicap[ensco$HomeTeam=="Rangers" & ensco$Date=="11/01/14"] <- -2.75
#Hamilton had a 12.5 goal handicap vs. Rangers; assume this meant 1.25:
ensco$HomeHandicap[ensco$HomeTeam=="Hamilton" & ensco$Date=="25/10/08"] <- 1.25
```

## Winning Probabilities

The correct probability (that is, the bookmaker consensus probability of the event that occurred) is found using a set of `for` loops: this is simple for the 1X2 market (using the `FTR` column) and the UO market (a new column, `TotGoals = FTHG + FTAG`, is created and assessed whether it is Under or Over 2.5), but requires more thought for the AH market.

First, the handicapped goal difference (coded as `gap`) between the two sides is found. Full-goal handicaps can result in a Home or Away Win, or a Draw (at which point the bet is considered void and the stake returned to the bettor (bet365, n.d.)). A half-goal handicap can only result in a Home or Away win. A quarter-goal handicap (E.G., $\frac{3}{4}$), however, can result in a *half-win* for the bettor: half of the stake is assigned to the nearest[1] half-handicap (in the case of $\frac{3}{4}$, this is $\frac{1}{2}$), and half to the nearest integer (1). If only one bet wins, the bettor wins on half their stake. The R code below is used for this step.

```
ensco$ah.gap <- with(ensco, FTHG.ah - FTAG); ensco$ah.res <- NULL
for (n in 1:N){
    if (ensco$ah.gap[n]<(-0.25)){ensco$ah.res[n]<-"aw"}
    else if (ensco$ah.gap[n]==(-0.25)){ensco$ah.res[n]<-"hfaw"}
    else if (ensco$ah.gap[n]==0){ensco$ah.res[n]<-"vo"}
    else if (ensco$ah.gap[n]==0.25){ensco$ah.res[n]<-"hfhm"}
    else if (ensco$ah.gap[n]>0.25){ensco$ah.res[n]<-"hm"}
    else{}
}
```

## Basic Calculations

The first step, as with Section 2.2, is to compute the bookmaker mean probabilities for each outcome, and their corresponding standard deviations, and compare against the observed probabilities. These are shown in Table 3.1, split by level.

To initially analyse this table, the 1X2 market is considered. as in Chapter 2, the standard deviation—for all levels—is far lower for Draws than for either Home or Away Wins, indicating a consistent lack of variation in the odds offered for Draws. In addition, the standard deviations for all three outcomes are lowest for Level 2 (rather than Level 3, as one would expect, due to the amount of, or lack thereof, information available at lower levels). For all three outcomes, across all three levels, the consensus probability is remarkably close to the observed probability; the clearest examples of this are the Level 2 Home Win probabilities: 0.4257 (consensus) and 0.4256 (observed), and the Level 1 Away Win probabilities: 0.2990 (consensus) and 0.2955 (observed).

[1] In a number-line sense

Table 3.1: Basic calculations for the entire English & Scottish data, forming part of the exploratory data analysis.

| **1X2 Market** | *Level 1* | *Level 2* | *Level 3* |
|---|---|---|---|
| Mean $\mathbb{P}_{\text{cons}}$(Home Win) | 0.4366 | 0.4257 | 0.4292 |
| Observed Probability of a Home Win | 0.4438 | 0.4256 | 0.4352 |
| Standard Deviation (Home Win 1x2) | 0.1507 | 0.1044 | 0.1238 |
| Mean $\mathbb{P}_{\text{cons}}$(Draw) | 0.2644 | 0.2725 | 0.2613 |
| Observed Probability of a Draw | 0.2607 | 0.2695 | 0.2406 |
| Standard Deviation (Draw 1x2) | 0.0369 | 0.0195 | 0.0243 |
| Mean $\mathbb{P}_{\text{cons}}$(Away Win) | 0.2990 | 0.3018 | 0.3095 |
| Observed Probability of a Away Win | 0.2955 | 0.3049 | 0.3242 |
| Standard Deviation (Away Win 1x2) | 0.1366 | 0.0952 | 0.1139 |

| **Under/Over 2.5 Goals Market** | *Level 1* | *Level 2* | *Level 3* |
|---|---|---|---|
| Mean $\mathbb{P}_{\text{cons}}$(Under 2.5 Goals) | 0.5080 | 0.5121 | 0.4756 |
| Observed Probability of Under 2.5 Goals | 0.5051 | 0.5184 | 0.4755 |
| Mean $\mathbb{P}_{\text{cons}}$(Over 2.5 Goals) | 0.4920 | 0.4879 | 0.5244 |
| Observed Probability of Over 2.5 Goals | 0.4949 | 0.4816 | 0.5245 |
| Standard Deviation (Under/Over Market) | 0.0568 | 0.0380 | 0.0504 |

| **Asian Handicap Market** | *Level 1* | *Level 2* | *Level 3* |
|---|---|---|---|
| Mean $\mathbb{P}_{\text{cons}}$(AH Home Win) | 0.5109 | 0.5046 | 0.5015 |
| Observed Probability of a Home Win (AH) | 0.4011 | 0.4005 | 0.4000 |
| Observed Probability of a Half-Home Win (AH) | 0.0434 | 0.0513 | 0.0583 |
| Mean $\mathbb{P}_{\text{cons}}$(AH Away Win) | 0.4891 | 0.4954 | 0.4985 |
| Observed Probability of an Away Win (AH) | 0.3779 | 0.3920 | 0.4087 |
| Observed Probability of a Half-Away Win (AH) | 0.0620 | 0.0741 | 0.0659 |
| Observed Probability of a Bet Being Voided | 0.1157 | 0.0821 | 0.0671 |
| Standard Deviation (Asian Handicap Market) | 0.0678 | 0.0524 | 0.046 |

For the UO and AH markets, bookmakers offer only two options, and thus, the standard deviations are equal for each outcome. For the UO market, the standard deviations are low (especially in Level 2, where $\sigma = 0.0380$); and the consensus and observed probabilities are very close together. In Level 3, the consensus and observed differ by 0.0001. The mean number of goals across all matches in the dataset is 2.66, so it is no surprise the means are around 0.5: it is likely the bookmakers know the mean number of goals in each level (or league), and set their odds accordingly.

Lastly, the AH market is considered. The numbers aren't as easy to infer across due to the half-wins and voided bets: of the full Home and Away Wins, 50.57% were Home Wins: by level, the proportions of full Home Wins are 0.5149, 0.5054 and 0.4946 for Levels 1, 2, and 3 respectively. Again, these are all remarkably close to the bookmaker consensus means. It is, as with the UO market, unsurprising these values are near 0.5: the AH market is designed to give a handicap in favour of the poorer side in the form of a goal deficit to the Home side (Constantinou, 2020).

## Visual Analysis

For visualisation of these values, several plots are produced. First are density plots, shown in Figure 3.1.

The densities of each outcome in the 1X2 market reflect similar distributions as in Figure 2.3. The Home Win curve is symmetrically distributed about the mean, located just above 0.4; the Away Win curve has a positive skew, with the mode between 0.25 and 0.3; and the Draw curve has a negative skew with a sharp drop after the mode, at roughly 0.27, with very little overall variance.

As expected, due to the nature of both markets, the UO and AH curves have little variation, with the mean for all four curves located at, or close to, 0.5. As both markets have two outcomes, the two curves for each market are reflections (Weisstein, n.d.). The last observation from these plots is the positive skew (and corresponding negative skew) of the Over 2.5 Goals curve (Under 2.5 Goals curve), suggesting the most common consensus probabilities have a slightly higher probability of Under 2.5 Goals, than Over.

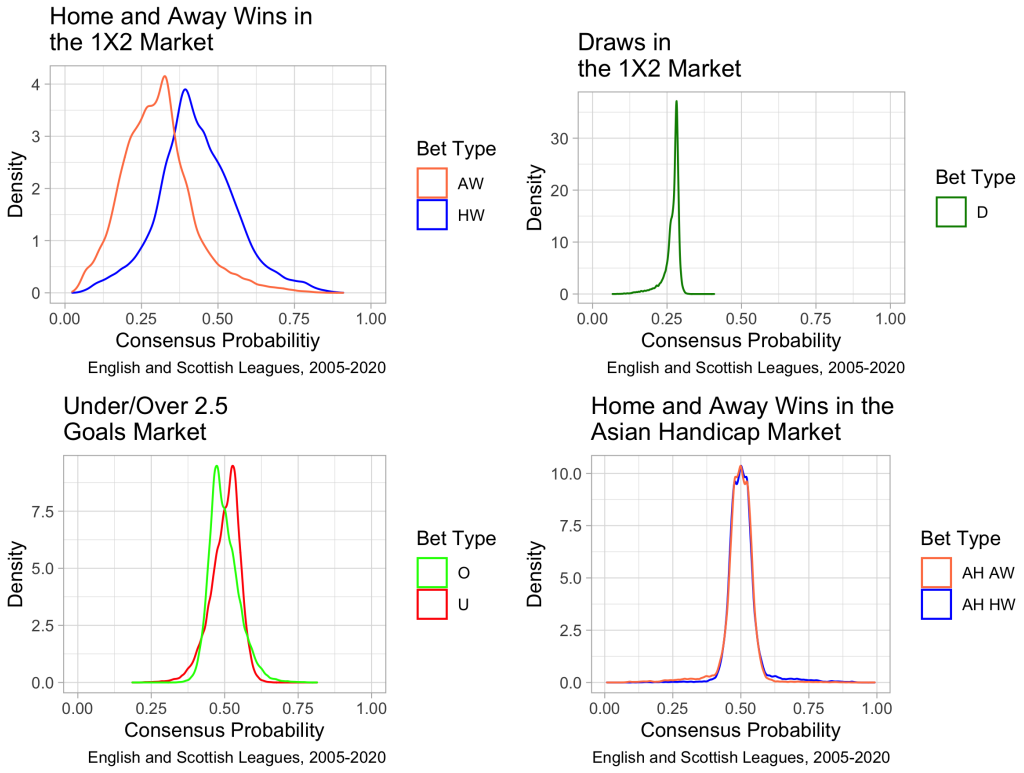


Figure 3.1: Density plots of the consensus probabilities offered in the 1X2, UO, and AH markets.

To assess the handicaps offered by bookmakers, two plots are created. The first, Figure 3.2: the handicap offered vs. consensus probability of a Home Win in the 1X2 market, with each data point grouped by level; and the second, Figure 3.3: the handicap offered vs. consensus probability of both a Home and Away Win in the 1X2 market. These plots are used to ensure the bookmakers handicap

is consistent with the consensus probabilities of a Home or Away Win without a handicap (the 1X2 market).

Both figures show that a team with a high consensus probability of a win, whether they are the Home or Away side, have a detrimental handicap: as expected. There are a number of outliers, perhaps due to errors made by `football-data.co.uk`, or by random chance.

There appears to be no difference in the handicap distribution between levels.

The non-linear shape may be due to small sample sizes (shown by the data points being reduced in size), with few matches at either extreme. In addition, across all levels, there are a number of matches with a 0 handicap, despite some of these matches being highly in favour of one side. This may be due to the data being unavailable, such as the Asian Handicap market being closed for that particular match.

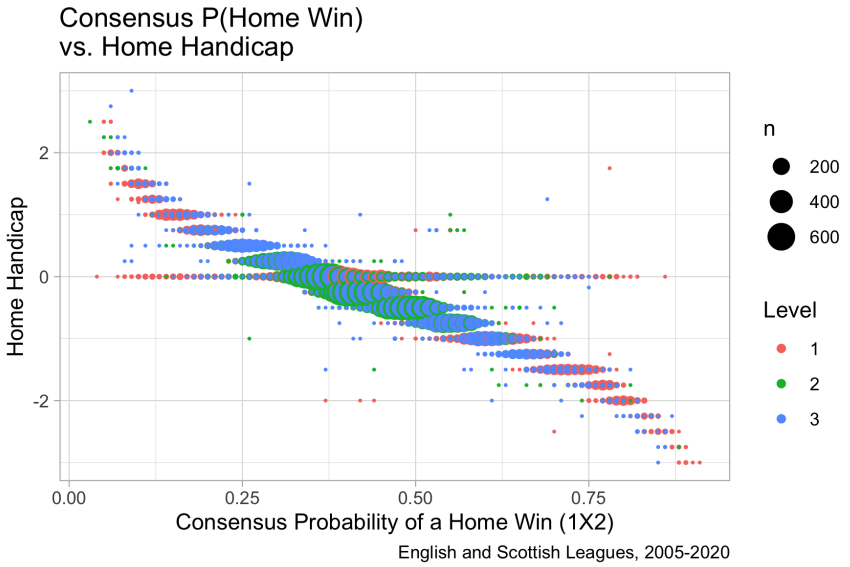


Figure 3.2: The AH handicap vs. the consensus probability of a Home Win, split by level.

Also created are tile plots, as shown before (Section 2.2, Figure 2.5). This is to see if the correct consensus probabilities are greater for more *convincing* wins, with both the 1X2 and UO markets considered. The bin size for each tile in both plots is given in Table 3.2.[2]

Figure 3.4, the tile plot for the 1X2 market, implies the expected result: more convincing results have a higher correct probability. The highest scoring tiles are highly-convincing Away Wins (0–5, 1–5, 2–5, 1–6+ and 0–3 are five examples of the most noticeable such tiles) and highly-convincing Home Wins (6+–0, 3–0 and 5–2), though the former is clearer. It is hard to draw meaningful conclusions from the highest scoring matches (the top right of the plot) due to low bin sizes, but at the bottom left (lowest scoring matches, where the number of matches per bin is high), the pattern holds well.

In Figure 3.5, the tile plot for the UO market, a different `ggplot2` palette is chosen, due to the low levels of variation in the market, and a black line is added at 2.5 Goals: the nine tiles within this

[2]The highest scoring Draw in the data for this chapter was 6–6, between Motherwell and Hibernian (05/05/2010) in the Scottish Premiership. 6+ goals is therefore chosen as the upper bound.

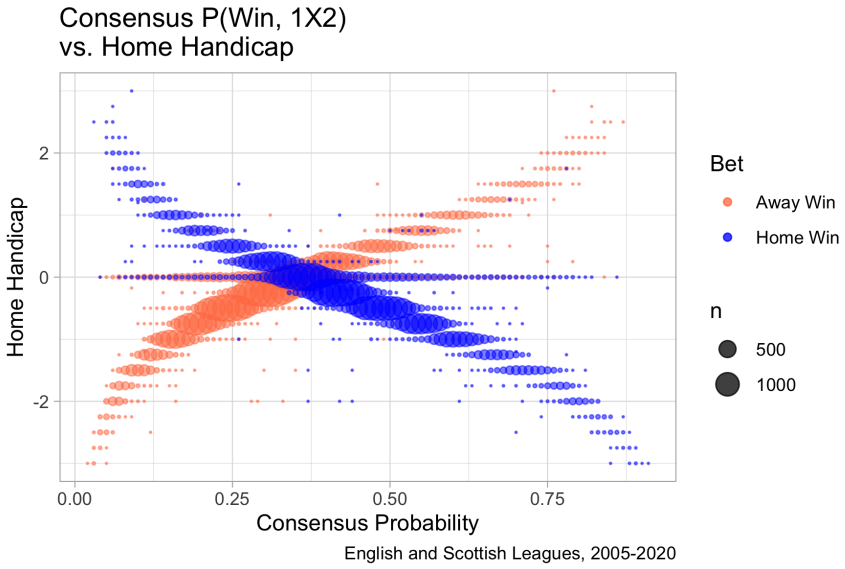


Figure 3.3: The AH handicap vs. the consensus probabilities of both a Home Win and Away Win.

bound (bottom left) contain Under 2.5 Goals; tiles outside contain Over 2.5 Goals. The 'redder' tiles are those with the highest consensus probability of Over 2.5 Goals[3] (1–5, 6+–1 and 0–3, with bin sizes $n_{1-5} = 133$, $n_{6+-1} = 119$ and $n_{0-3} = 1085$); 'bluer' tiles are those with the lowest consensus probability of Over 2.5 Goals. The 5–3 Home Win tile ($n_{5-3} = 47$) has the lowest correct probability, at around 0.4, despite eight goals being scored in these matches. Other high-scoring tiles with low correct probabilities include the 1–4, 2–6+, and 6+–6+. Each of these has a low number of matches, $n_{1-4} = 446$, $n_{2-6+} = 16$, $n_{6+-6+} = 1$. The most striking conclusion this tile plot gives, however, is the *lack* of a pattern: there is no obvious feature, suggesting—as with the Draws in the 1X2 market—bookmakers struggle at placing reliable odds in the UO market.

Table 3.2: The bin size for each tile of Figures 3.4 and 3.5.

| Away Goals | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 6+ | 48 | 38 | 16 | 12 | 1 | 1 | 1 |
| | 5 | 115 | 133 | 74 | 23 | 18 | 6 | — |
| | 4 | 407 | 446 | 268 | 144 | 54 | 19 | 5 |
| | 3 | 1085 | 1337 | 1003 | 508 | 167 | 47 | 7 |
| | 2 | 2388 | 3530 | 2628 | 1242 | 411 | 125 | 44 |
| | 1 | 4091 | 6012 | 4372 | 1993 | 763 | 245 | 119 |
| | 0 | 3590 | 5009 | 3761 | 1954 | 782 | 304 | 132 |
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6+ |
| | | Home Goals | | | | | | |

[3]The highest correct consensus probability, although all such tiles lie within the Over 2.5 Goals region.

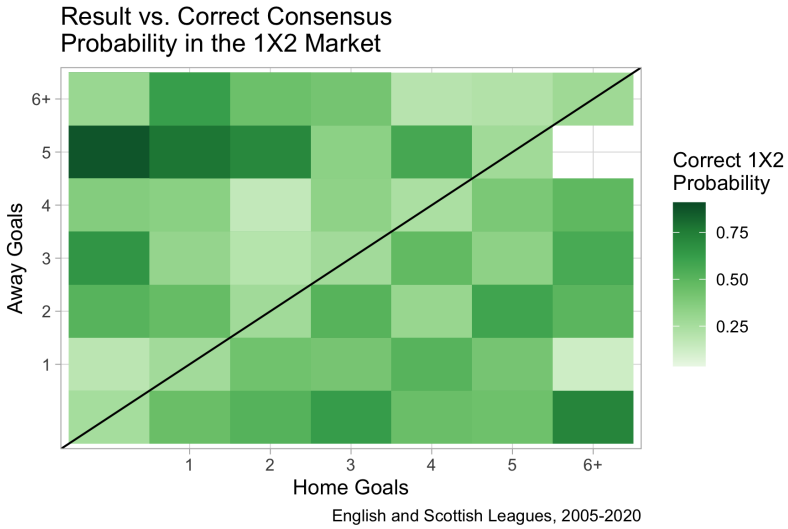


Figure 3.4: Tile plot of the correct 1X2 probability vs. the full-time result.

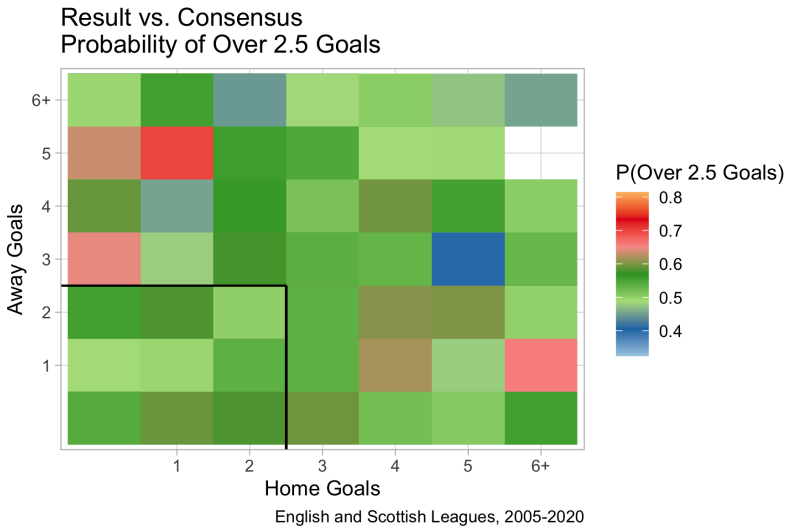


Figure 3.5: Tile plot of the correct UO probability vs. the full-time result.

The final two exploratory plots created are *count plots*. These are a way of visualising two DISCRETE variables, with larger points indicating more data with the corresponding values (Plotly.com, n.d.). The first, Figure 3.6, is a plot of the consensus home handicap offered for a match against the consensus probability of Over 2.5 Goals.[4] It would be expected that matches with a larger handicap in magnitude (bookmakers estimate these will be the most convincing matches) will have a larger

[4]Rounded to 2 decimal places and considered discrete.

consensus probability of Over 2.5 Goals. The latter, Figure 3.7, is a plot of the expected vs. actual goal difference, where the expected is the handicap, the actual is `FTAG` − `FTHG`. If the bookmakers are setting accurate handicaps, this would be expected to a linear relationship.

In both plots, the expected general trend is observed, and it can assumed, based on these two plots and Figures 3.2 and 3.3, that the handicaps set by bookmakers are well set. The accuracy of the markets requires further investigation.

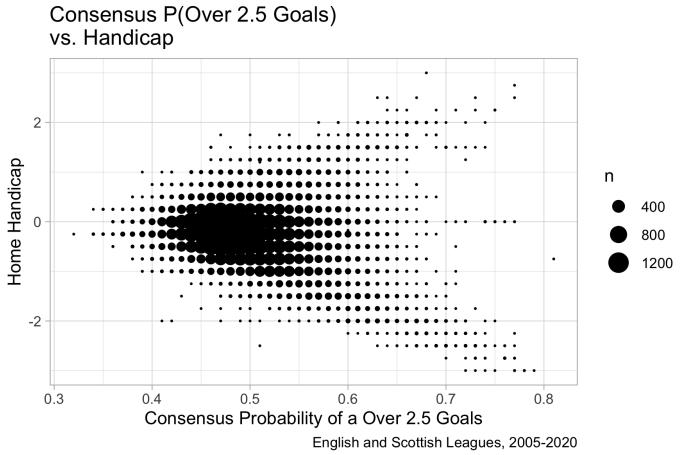


Figure 3.6: Count plot of the home handicap offered vs. the consensus probability of over 2.5 goals.

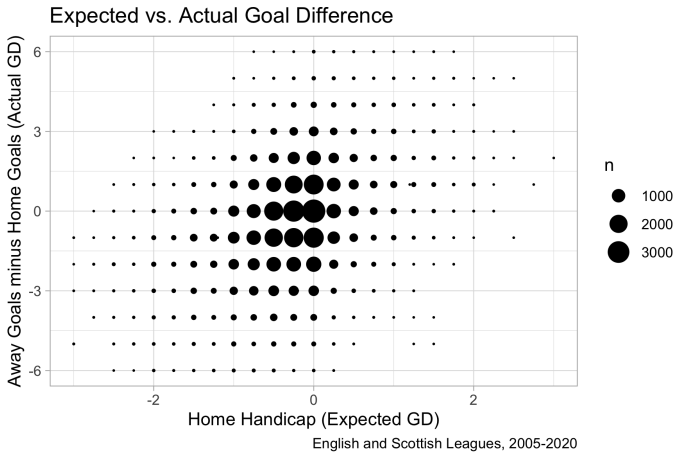


Figure 3.7: Count plot of the expected vs. actual goal difference.

## 3.2 Correlation Analysis

A correlation analysis is conducted as in Chapter 2. For all three markets, the data is first binned ($n = 50$ is chosen across all markets), done in R using the `cut` and `tapply` commands. The observed probabilities, as defined previously as being the observed proportion of matches in a bin with the outcome of interest, of each bin are found using the `prop.table` function. For the Asian Handicap market, only full wins are considered for ease. As the Under/Over 2.5 Goals and Asian Handicap markets are binary, only one linear model is created. The coefficients of the six final models to predict the observed probability $\mathfrak{O}$ from the consensus probability $\mathfrak{C}$ are shown in Table 3.3. For 'ideal' models, the gradient, or *slope* of the linear model would be equal to 1.

Table 3.3: Linear models to predict the observed outcome $\mathfrak{O}$ from a given bookmaker consensus probability $\mathfrak{C}$ using the English & Scottish data.

| *Market* | *Intercept* | *Slope* |
|---|---|---|
| 1X2 Home Win | -0.0153 | 1.0637 |
| 1X2 Draw | -0.1797 | 1.9814 |
| 1X2 Away Win | -0.0111 | 1.0462 |
| 1X2 Overall | -0.0089 | 1.0642 |
| Under/Over | 0.0358 | 0.8463 |
| Asian Handicap | -0.0419 | 0.8691 |

These models can be used for predictions, for example, if one was to know the consensus probabilities for an Asian Handicap bet and a Home Win 1X2 bet were both 0.6, it can be predicted that the respective observed probabilities will be 0.48 and 0.62. The models aren't much help for this use: instead, they are used to find values for the coefficient of determination $R^2$ and root-mean-square-error RMSE, and to examine the slope; these values are in Table 3.4.

Table 3.4: Values for $R^2$, RMSE, and Slope for the markets of interest, based on the models in Table 3.3.

| | **1X2** | | |
|---|---|---|---|
| | *Home Win* | *Draw* | *Away Win* |
| $R^2$ | 0.9917 | 0.5177 | 0.97148 |
| RMSE | 0.0247 | 0.1602 | 0.04393 |
| Slope | 1.0637 | 1.9814 | 1.0462 |
| | **Under/Over** | **Asian Handicap** | |
| $R^2$ | 0.3868 | 0.8959 | |
| RMSE | 0.1478 | 0.0756 | |
| Slope | 0.8463 | 0.8691 | |

The values clearly indicate poor levels of accuracy in the UO market, with $R^2 = 38.7\%$, RMSE $= 0.1478$—only lower than the RMSE for Draws in the 1X2 market—and a slope of 0.85. High levels of accuracy are found in the AH market (low RMSE, and $R^2 \approx 90\%$), though the slope is below the ideal, at 0.87. The 1X2 market findings are consistent with the findings outlined in Chapter 2, with low levels of accuracy for Draws; high levels of accuracy for Home and Away Wins, both across all

three measures. To visualise these findings, plots are produced. These are shown in Figures 3.8, 3.9, and 3.10. The figures also contain the linear model and corresponding 95% confidence interval, as defined in Appendix A.1, Definition 5.

The figures support the conclusions from the above values. Figure 3.8 has a near-perfect correlation for the Home Wins and Away Wins, corresponding to the respective $R^2$ values: 0.992 and 0.971, with the slope closely aligned to the ideal. The Draws are inaccurate, with a large confidence interval, implying a large standard error (in fact, the SE is 0.3063). In addition, the slope is visually inaccurate, corresponding to the value of 1.98. Figure 3.9 suggests relatively high levels of accuracy in the UO market; the $R^2$ and RMSE values may be due to high values of leverage at the extremities, emphasised by the low levels of variance (Table 3.1) in the odds offered. Figure 3.10 also shows high levels of accuracy in the AH market, though both lines are significantly[5] below the $y = x$ (dotted) line, possibly due to bookmaker overround. To further analyse these values, each level is considered separately.

[5]No part of the 95% confidence interval lies above the dotted line.

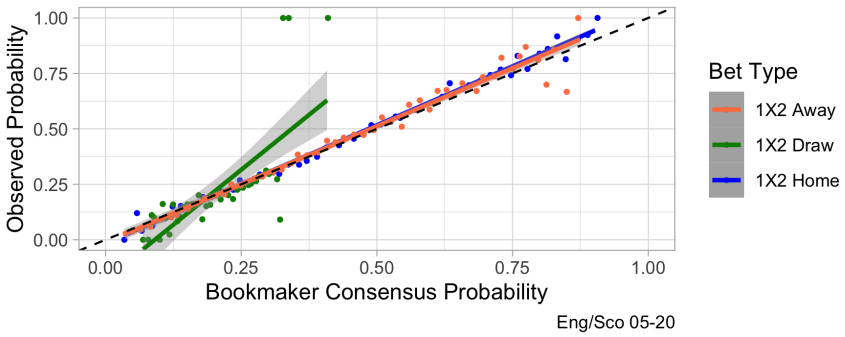


Figure 3.8: Consensus vs. observed probabilities on the English & Scottish data, 1X2 market.

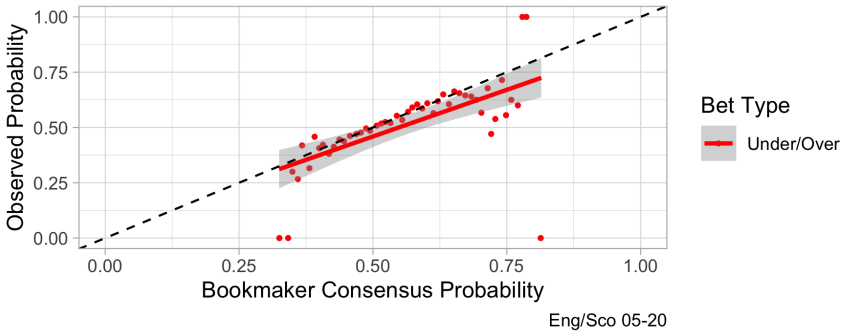


Figure 3.9: Consensus vs. observed probabilities on the English & Scottish data, Under/Over 2.5 Goals market.

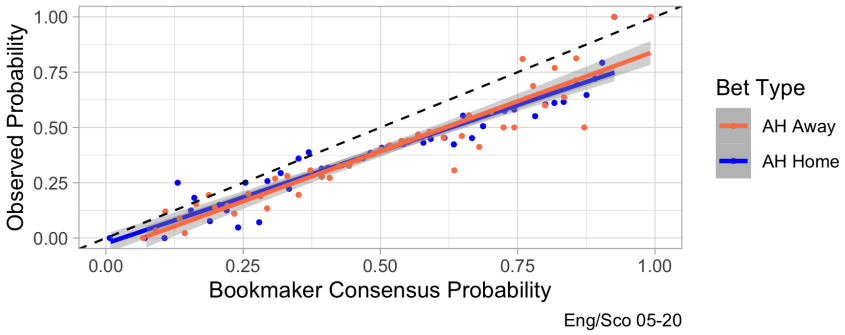


Figure 3.10: Consensus vs. observed probabilities on the English & Scottish data, Asian Handicap market.

## 3.3 Comparing Levels

In addition to finding $R^2$, RMSE, and the slope, the values of $P_1$ and $P_2$ are also found for each market. As in Section 2.5, in order to find the values from correlation analysis, the bins are assigned a weight, with more weight given to the Home Win and Away Win bets (the matches are split into 35 bins) than to the less-varied Draw bets (15 bins). For UO and AH markets, the number of bins is 35. Plots for each are shown in Figures 3.11, 3.12, and 3.13. The values of statistical accuracy are provided in Table 3.5.

Table 3.5: $R^2$, RMSE, $P_1$, $P_2$, and Slope values for the 1X2, UO, and AH markets across all three levels.

| | *Level 1* | *Level 2* | *Level 3* |
|---|---|---|---|
| Number of matches $n$ | 17317 | 18913 | 13248 |
| **1X2 Market** | | | |
| $R^2$ | 0.99035 | 0.75579 | 0.91903 |
| RMSE | 0.02611 | 0.12914 | 0.07094 |
| $P_1$ | 0.36767 | 0.35137 | 0.35902 |
| $P_2$ | 0.59823 | 0.62984 | 0.61429 |
| Slope | 1.08106 | 0.95331 | 1.04732 |
| **Under/Over 2.5 Goals Market** | | | |
| $R^2$ | 0.63719 | 0.53157 | 0.02919 |
| RMSE | 0.10474 | 0.12414 | 0.12980 |
| $P_1$ | 0.50331 | 0.50129 | 0.50271 |
| $P_2$ | 0.49343 | 0.49743 | 0.49458 |
| Slope | 1.03996 | 1.14681 | 0.18112 |
| **Asian Handicap Market** | | | |
| $R^2$ | 0.81714 | 0.76950 | 0.46815 |
| RMSE | 0.09766 | 0.10212 | 0.19181 |
| $P_1$ | 0.50582 | 0.50374 | 0.50238 |
| $P_2$ | 0.48920 | 0.49277 | 0.49534 |
| Slope | 0.82633 | 0.86569 | 0.76894 |

Both the figures and values suggest, in addition to having less variation, bookmaker performance is worse in Level 2 than in Level 3 in the 1X2 market, suggested by four metrics. (The slope is roughly 4.7% away from the ideal in both, though Level 2 lies below the ideal and Level 3 is above.) Visually, the confidence intervals are larger for Level 2 than Level 3, too. Whilst the number of games in Level 3 is less ($n_{L3} = 13248 < 18913 = n_{L2}$), it is unlikely to be reason, instead being down to reasons such as the favourite-longshot bias, competitive balance,[6] and/or simply bookmakers struggling more to place accurate odds on matches at Level 2 than Level 3.

In the UO market, whilst the $P_1$ values for all three levels are higher, and the $P_2$ lower, than in the 1X2 market, the two markets cannot be compared with each other this way, as there are less options to bet on. Between levels, Level 1 has the highest $P_1$ and lowest $P_2$, indicating—as with

[6] In Level 2, all teams are professional, whereas in Level 3, there are semi-professional teams—who have have been recently promoted from regional leagues—competing against fully professional teams—such as those recently relegated—meaning the 'gulf' in quality between the two may be larger between the best and worse teams in Level 3 than 2.

the 1X2 market—better performance at the highest level. This is shown with the $R^2$ and RMSE values, too. The $P$-values suggest better predictive performance in Level 3 than 2, with a higher $P_1$ and lower $P_2$, and a similar RMSE value. This is not, however, replicated with the coefficient of determination, with Level 2 having an $R^2$ of 53%, much higher than the 3% of Level 3. The $R^2$ and RMSE across all three markets indicate poor performance. The slope is accurate in Level 1, only 0.4 from the ideal; in Level 2, this grows to 0.15, and in Level 3, this is over 0.8 below the ideal line, the worst slope across all nine market-level combinations.

The AH market has a vastly lower $R^2$ value in Level 3, 46.8%, than Levels 1 and 2, at 82% and 77% respectively. In addition, the $P_1$ values descend down the levels, and $P_2$ ascend, indicating worse performance at lower levels. Compared to the UO market, in each level, the $P_1$ is lower and $P_2$ higher for the AH market, indicating a better performance. It is worth noting that the overall $R^2$ of the AH market was shown to be 90%, higher than all three levels. This is due to the problem of $R^2$ being reliant on the number of bins used in an analysis, though the general trend is of interest here. The values show a decrease in accuracy as lower levels are considered, with Level 3 having the lowest $R^2$, highest RMSE, and the slope furthest from the ideal. The slope is, however, more accurate in Level 2 than 1.

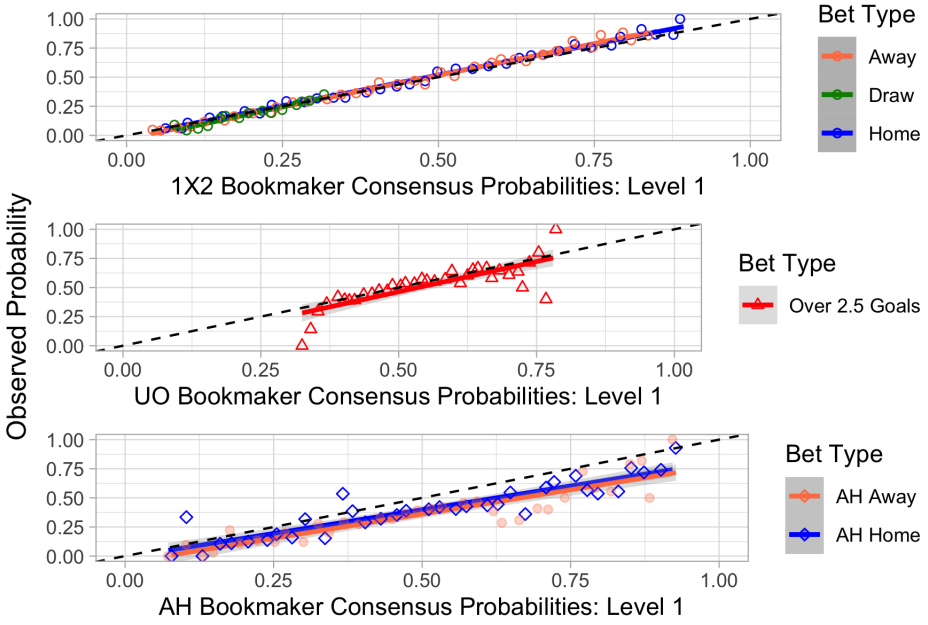


Figure 3.11: Plots of consensus vs. observed probabilities, English & Scottish data, Level 1.

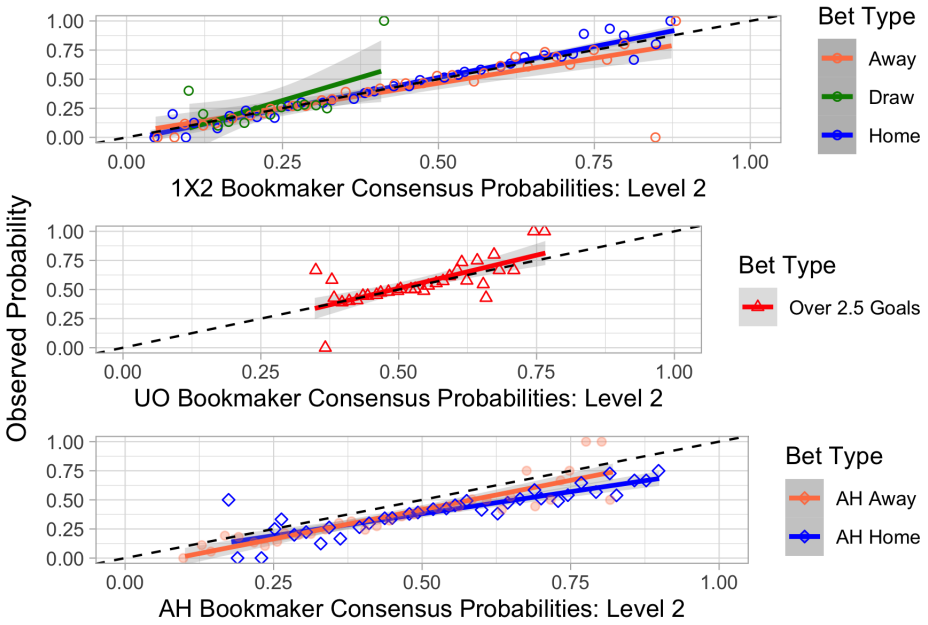


Figure 3.12: Plots of consensus vs. observed probabilities, English & Scottish data, Level 2.

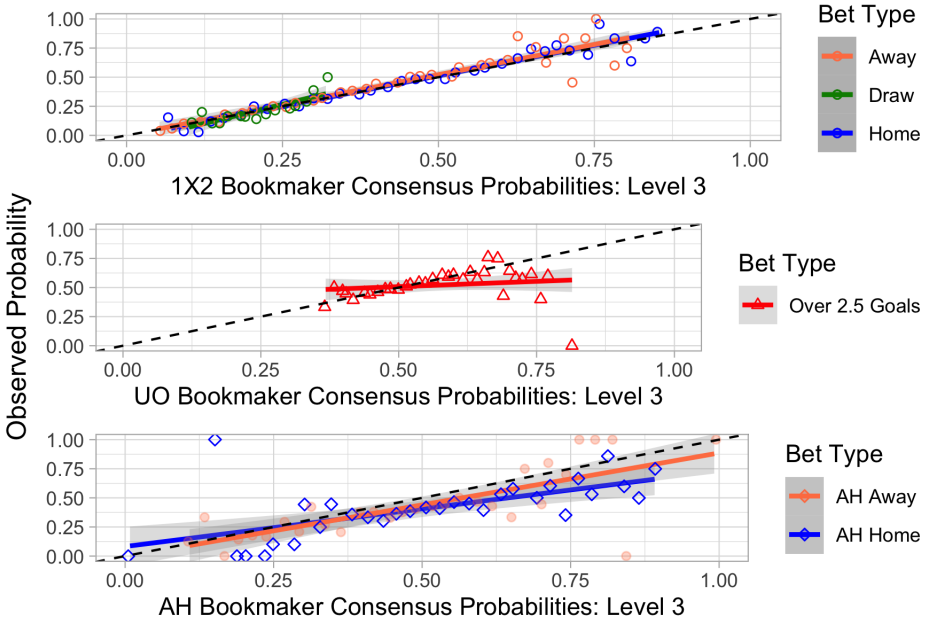


Figure 3.13: Plots of consensus vs. observed probabilities, English & Scottish data, Level 3.

## 3.4 Comparing Seasons

As with Section 2.7, the $R^2$, RMSE, $P_1$, $P_2$, and Slope values are found for each season in the 1X2 market, as well as the same five values for the UO and AH markets. These are given in Table 3.6, and displayed in Figure 3.14, with each market's plots in a column. For the slope plots, on the fifth row, the ideal line, where the slope is equal to 1, is added in black.

In the 1X2 market, unlike as seen previously in Section 2.7, the lines of best fit for the $R^2$ and RMSE are not improving over time. It is likely, in both cases, due to two outliers: the 15/16 season ($R^2 = 87\%$, RMSE $= 0.0736$) and the 17/18 season ($R^2 = 78\%$, RMSE $= 0.1200$), as opposed to an actual trend, which appears to be stable from the remaining points. The $P_1$ and $P_2$ values and the slope (which is approaching the black, ideal line at $y = 1$) for the 1X2 market, however, indicate improvement over time.

The UO market shows great improvement by the $R^2$ and RMSE metrics over time, with clear increasing and decreasing trends respectively. In addition, the $P_1$ values are increasing, and $P_2$ decreasing, against suggesting improvements. The slope is approaching the ideal, too. The UO market was found to be inaccurate in Section 3.1, though if this trend continues, this finding may be out-of-date soon. This could be due to improvements in modelling and new metrics/statistics such as 'Expected Goals' (Rathke, 2017; footballxg.com, n.d.).

The AH, however, shows a deterioration over time across $R^2$ (decreasing) and RMSE (increasing), with the former going from a peak of 97% in the 09/10 season to a low of 0.04% in the 14/15 season. A potential reason for this is the decrease in variation of consensus probabilities offered in this market, which is investigated in Section 3.5, Figure 3.20. The same pattern is shown with the $P$-values, with $P_1$ decreasing and $P_2$ increasing, though both lines of best fit are skewed by the 2009/10 outlier (with the highest-ranking performance in three of the four metrics).

Table 3.6: Statistics for accuracy across each market, by season.

| | **1X2 Market** | | | | | **UO Market** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Season* | $R^2$ | RMSE | $P_1$ | $P_2$ | Slope | $R^2$ | RMSE | $P_1$ | $P_2$ | Slope |
| 05/06 | 0.9728 | 0.0389 | 0.3734 | 0.5882 | 1.1119 | 0.0142 | 0.1653 | 0.5020 | 0.4960 | -0.2399 |
| 06/07 | 0.9605 | 0.0413 | 0.3680 | 0.5985 | 1.0324 | 0.0015 | 0.1285 | 0.5019 | 0.4962 | 0.0767 |
| 07/08 | 0.9650 | 0.0523 | 0.3708 | 0.5931 | 1.2420 | 0.7222 | 0.0906 | 0.5011 | 0.4977 | 1.8784 |
| 08/09 | 0.9837 | 0.0295 | 0.3750 | 0.5852 | 1.1483 | 0.0162 | 0.2418 | 0.5017 | 0.4966 | -0.2598 |
| 09/10 | 0.9750 | 0.0377 | 0.3807 | 0.5752 | 1.1136 | 0.0348 | 0.2050 | 0.5014 | 0.4972 | 0.3705 |
| 10/11 | 0.9634 | 0.0486 | 0.3691 | 0.5960 | 1.1669 | 0.0221 | 0.0932 | 0.5018 | 0.4964 | 0.1640 |
| 11/12 | 0.9773 | 0.0347 | 0.3763 | 0.5829 | 1.0868 | 0.8098 | 0.0352 | 0.5021 | 0.4958 | 0.7772 |
| 12/13 | 0.9677 | 0.0392 | 0.3792 | 0.5762 | 0.9640 | 0.4759 | 0.0876 | 0.5034 | 0.4933 | 0.8044 |
| 13/14 | 0.9731 | 0.0446 | 0.3864 | 0.5629 | 1.2052 | 0.7314 | 0.0586 | 0.5052 | 0.4896 | 0.8480 |
| 14/15 | 0.9804 | 0.0348 | 0.3799 | 0.5756 | 1.1579 | 0.1705 | 0.0663 | 0.5032 | 0.4936 | 0.3018 |
| 15/16 | 0.8714 | 0.0736 | 0.3773 | 0.5806 | 0.8367 | 0.8141 | 0.0453 | 0.5028 | 0.4945 | 0.8132 |
| 16/17 | 0.9639 | 0.0511 | 0.3927 | 0.5528 | 1.1608 | 0.0240 | 0.1380 | 0.5027 | 0.4945 | 0.1923 |
| 17/18 | 0.7835 | 0.1200 | 0.3920 | 0.5536 | 1.0335 | 0.8656 | 0.0570 | 0.5013 | 0.4975 | 1.1643 |
| 18/19 | 0.9740 | 0.0432 | 0.3851 | 0.5651 | 1.1056 | 0.7915 | 0.0374 | 0.5015 | 0.4969 | 0.6041 |
| 19/20 | 0.9676 | 0.0433 | 0.3772 | 0.5797 | 0.9888 | 0.9255 | 0.0390 | 0.5039 | 0.4923 | 1.1233 |

| | **AH Market** | | | | |
|---|---|---|---|---|---|
| *Season* | $R^2$ | RMSE | $P_1$ | $P_2$ | Slope |
| 05/06 | 0.8091 | 0.1568 | 0.5026 | 0.4948 | 1.9069 |
| 06/07 | 0.7984 | 0.1002 | 0.5031 | 0.4942 | 0.9581 |
| 07/08 | 0.8495 | 0.0977 | 0.5053 | 0.4900 | 0.9842 |
| 08/09 | 0.8538 | 0.0880 | 0.5044 | 0.4916 | 0.7990 |
| 09/10 | 0.9725 | 0.0335 | 0.5290 | 0.4480 | 0.8402 |
| 10/11 | 0.6364 | 0.1217 | 0.5059 | 0.4884 | 0.7321 |
| 11/12 | 0.3550 | 0.1624 | 0.5023 | 0.4954 | 1.2801 |
| 12/13 | 0.0063 | 0.2176 | 0.5018 | 0.4964 | 0.0969 |
| 13/14 | 0.0033 | 0.2742 | 0.5022 | 0.4955 | 0.0962 |
| 14/15 | 0.0004 | 0.0693 | 0.5005 | 0.4990 | -0.0194 |
| 15/16 | 0.5281 | 0.1871 | 0.5023 | 0.4954 | 1.3587 |
| 16/17 | 0.1074 | 0.0713 | 0.5006 | 0.4988 | 0.3846 |
| 17/18 | 0.1340 | 0.1946 | 0.5012 | 0.4975 | 0.9118 |
| 18/19 | 0.2352 | 0.1614 | 0.5010 | 0.4981 | 0.9601 |
| 19/20 | 0.5271 | 0.1605 | 0.5006 | 0.4988 | 1.7778 |

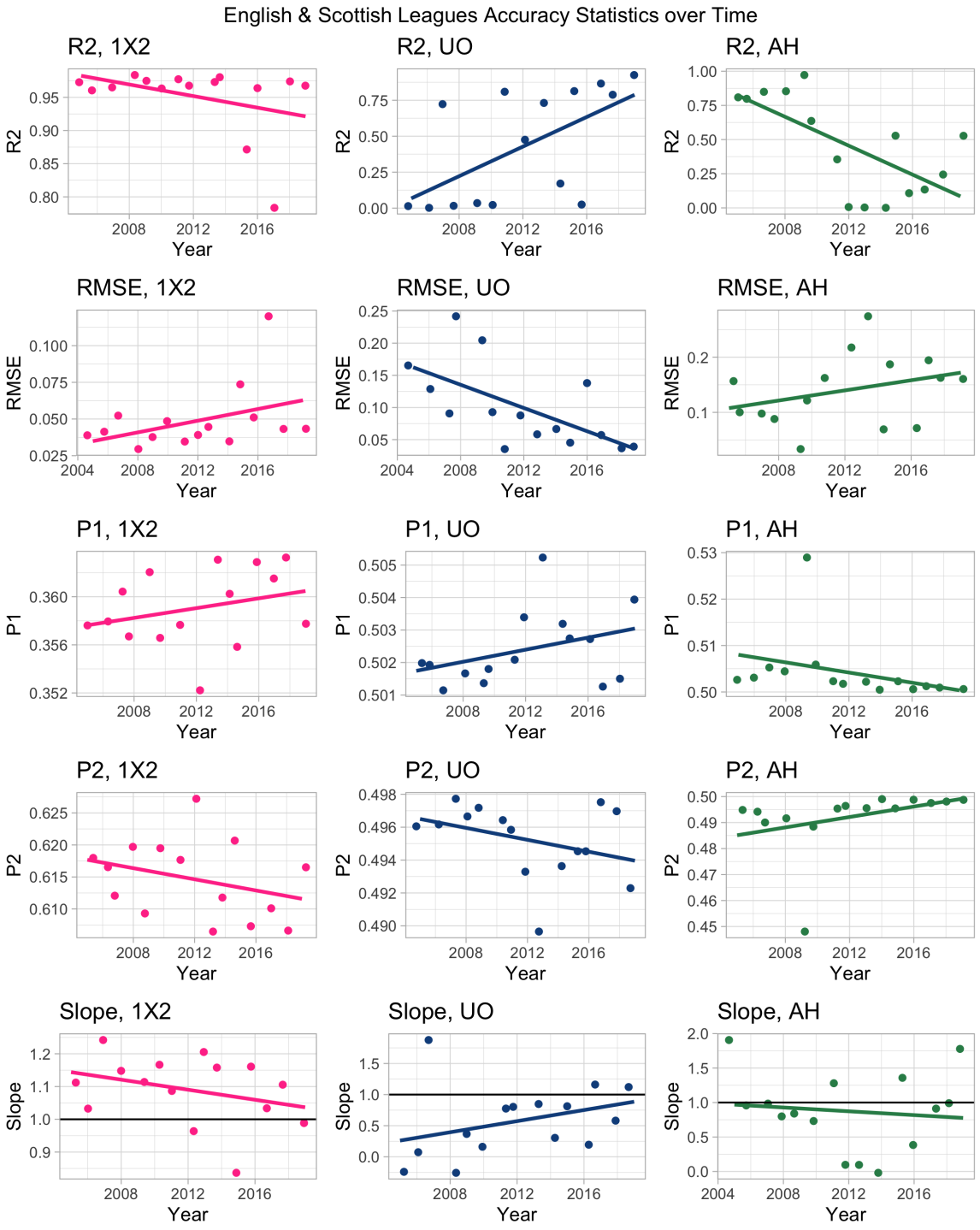


Figure 3.14: The variation of $R^2$, RMSE, $P_1$, $P_2$, and the Slope over time in the English & Scottish leagues data

## 3.5 The Overround

The *overround* $\eta$ for a single match in a market with $k$ outcomes is defined in Equation 3.1 (MatterOfStats, n.d.). An overround of 1.05 (the sum of underlying probabilities from the odds = 1.05) represents a 5% bookmaker commission for that match, in that market. The overround is not constant, either, and may change among 'matches, bookmakers and over time' (Angelini and De Angelis, 2019), as well as being disproportionate across each outcome in a single market, with bookmakers taking advantage of bettor's biases (Levitt, 2004). As such, an assessment into the outcomes (E.G., 1X2 Home Win) cannot be conducted, only the markets (E.G., 1X2) as a whole. To assess where the bookmakers earn their commission, calculations and jitter/scatter plots of the overround are considered in a given market, grouped by certain variables (level and season). The values are in Table 3.7; the plots are shown in Figures 3.15 to 3.20.

$$\eta = \sum_{i=1}^{k} \left[ \frac{1}{O_i} \right] = \sum_{i=1}^{k} [\mathbb{P}_{\text{und}}(i)] \tag{3.1}$$

Table 3.7: The mean overround $\bar{\eta}$ for different groups of the data.

| | **1X2 Market** | | **UO Market** | | **AH Market** | |
|---|---|---|---|---|---|---|
| Overall $\bar{\eta}$ | 1.0825 | 8.25% | 1.0684 | 6.84% | 1.0470 | 4.70% |
| Level 1 $\bar{\eta}$ | 1.0707 | 7.07% | 1.0662 | 6.62% | 1.0461 | 4.61% |
| Level 2 $\bar{\eta}$ | 1.0838 | 8.38% | 1.0685 | 6.85% | 1.0473 | 4.73% |
| Level 3 $\bar{\eta}$ | 1.0961 | 9.61% | 1.0712 | 7.12% | 1.0477 | 4.77% |
| Season 05/06 $\bar{\eta}$ | 1.1110 | 11.10% | 1.0798 | 7.89% | 1.0436 | 4.36% |
| Season 12/13 $\bar{\eta}$ | 1.0769 | 7.69% | 1.0635 | 6.35% | 1.0400 | 4.00% |
| Season 19/20 $\bar{\eta}$ | 1.0659 | 6.59% | 1.0592 | 5.92% | 1.0468 | 4.69% |

### Overround by Level

Figures 3.15, 3.17, and 3.19 show the overround of each match in the 1X2, UO, and AH markets respectively, grouped by level (with red, green, and blue points for Levels 1, 2, and 3). In the 1X2 market, Level 1 has more varied overround, ranging from around 3% to 11%; Level 3 varies from approximately 7% to 12%.[7]

In the UO market, the overround seems to be more evenly distributed, with more variation (across all levels) for matches when the odds are nearer 2 ($\mathbb{P}_{\text{cons}} = 0.5$), forming an almost symmetric bell curve shape. In this market, the overround is generally lower than for the 1X2 market, varying from just around 2.5% to just over 12%, with the majority around 5% to 7%. Level 1 odds vary less than Levels 2 and 3, too.

The AH market has a less varied overround. The majority of matches exist between 5% and 10%; when the consensus probability is near 0.5, however, the overround appears to decrease to a low of around 1%. This pattern holds across all three levels.

[7] It is worth noting the matches are plotted first in order of time, then in order of league, so more recent matches in Level 3 are the datapoints added last. This can be partially resolved by altering the `alpha` (opacity) of the points. This is chosen to be 50%.

These findings are backed up by the values in Table 3.7. In the 1X2 market, the mean overround in Level 1 is 7.1%, whereas in Level 3, this rises to 9.6%. In the UO market, it rises from 6.6% in Level 1 to 7.1% in Level 3; for the AH market, it is more stable, rising from 4.6% to 4.8%.

## Overround by Season

The shape of these figures will be identical to those for overround by Level in their respective markets. Figures 3.16, 3.18, and 3.20 show the overround of each match, grouped by season, with the colour of the data points following the colours of a rainbow: reds and yellows are for earlier seasons, greens and blues for 'middle' seasons, and violets and pinks for later (more-recent) seasons.

In the 1X2 market 3.16, overround is decreasing over time, with the maximum overround decreasing from around 12% for the earliest seasons in the data to around 10%. The floor of the range hasn't moved from around 4%, however.

In the UO market, the variation in the overround has decreased massively: the pink datapoints for the 19/20 season cover a range from 5% to 7%. At the bulge when $\mathbb{P}_{\text{cons}} \approx 0.5$, the matches that lie above or below this range are generally from earlier seasons, varying from around 2.5% to just over 12%.

From Figure 3.20, the most striking change is in the variation in consensus probabilities: these range from around 0.12 to 0.80 for the 'middle'/green datapoints, reducing to a high concentration around 0.4 to 0.6 in more recent (pink) seasons, indicating as time progresses, bookmaker handicap selection is improving. The overround is slightly decreased too in the 'bulge' with this high-concentration area having an overround around 3% to 6% in recent seasons and 2% to 10% in older seasons; the green band from 0.12 to 0.80 has a higher overround, around 6% to 10%.

The values in Table 3.7 again back this up. Over time, the mean 1X2 overround has decreased from 11.1% to 6.6% and the mean UO overround from 8.0% to 5.9%. The mean AH overround has changed from 4.4% to 4.7%, despite being as low as 4.0% in 2012/13 (half-way through the dataset).

## Summary of Findings

To summarise, the overround is reduced at higher levels, across all three markets: this is most evident in the 1X2 market (where the mean Level 3 overround is 2.54% higher than Level 1), and is least evident in the Asian Handicap market (0.16%). Over time, the overround has been reducing the 1X2 and Under/Over 2.5 Goals markets, but remaining stable in the Asian Handicap: the largest change in the AH market is the variance of the consensus probability (and thus, odds offered), which have been reducing over time.

## Overround Plots

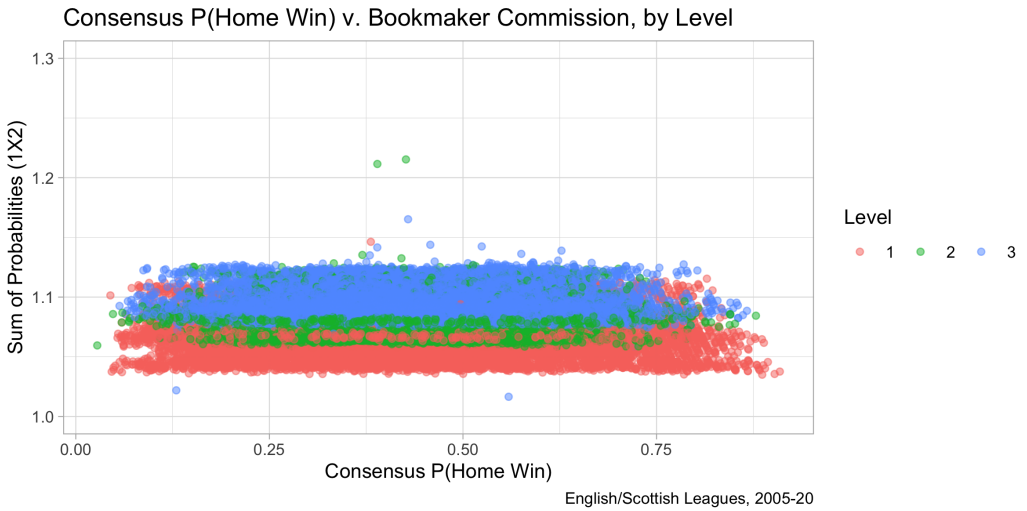


Figure 3.15: The overround in the 1X2 market vs. $\mathbb{P}_{\text{cons}}$(1X2 Home Win), split by level.

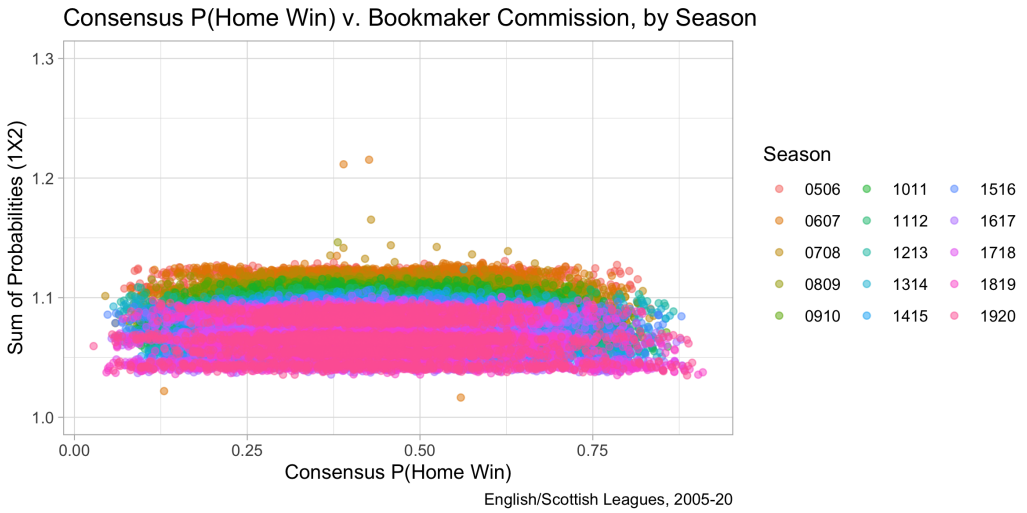


Figure 3.16: The overround in the 1X2 market vs. $\mathbb{P}_{\text{cons}}$(1X2 Home Win), split by season.

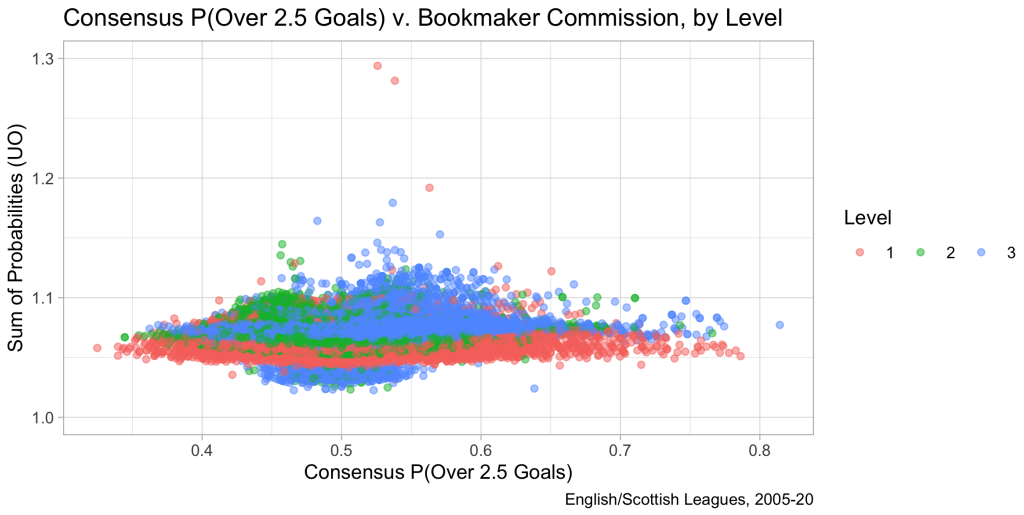


Figure 3.17: The overround in the UO market vs. $\mathbb{P}_{\text{cons}}$(Over 2.5 Goals), split by level.

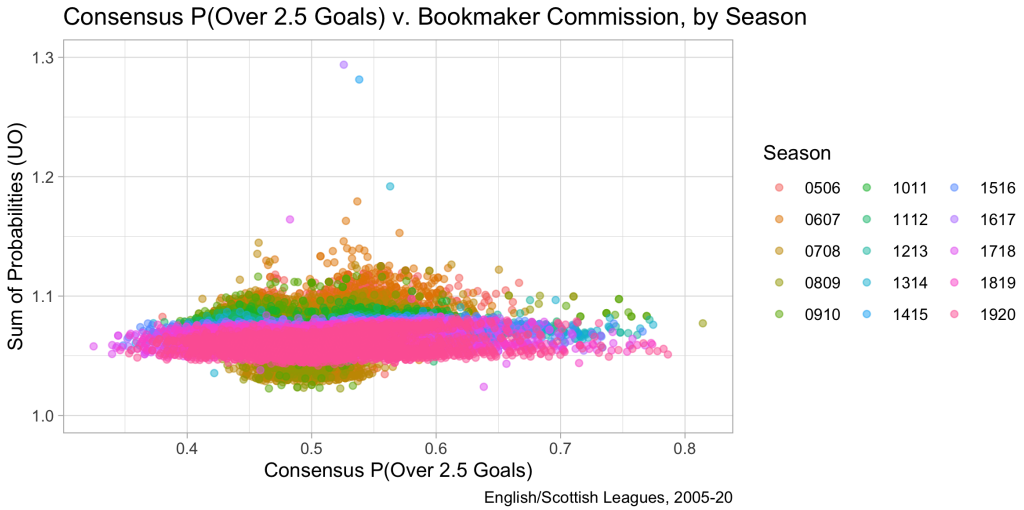


Figure 3.18: The overround in the UO market vs. $\mathbb{P}_{\text{cons}}$(Over 2.5 Goals), split by season.

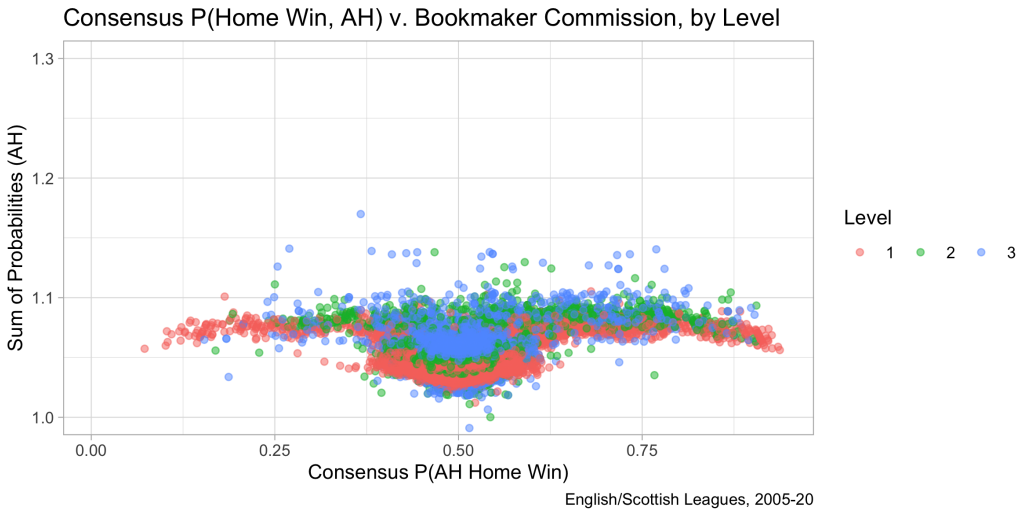


Figure 3.19: The overround in the AH market vs. $\mathbb{P}_{\text{cons}}$(AH Home Win), split by level.

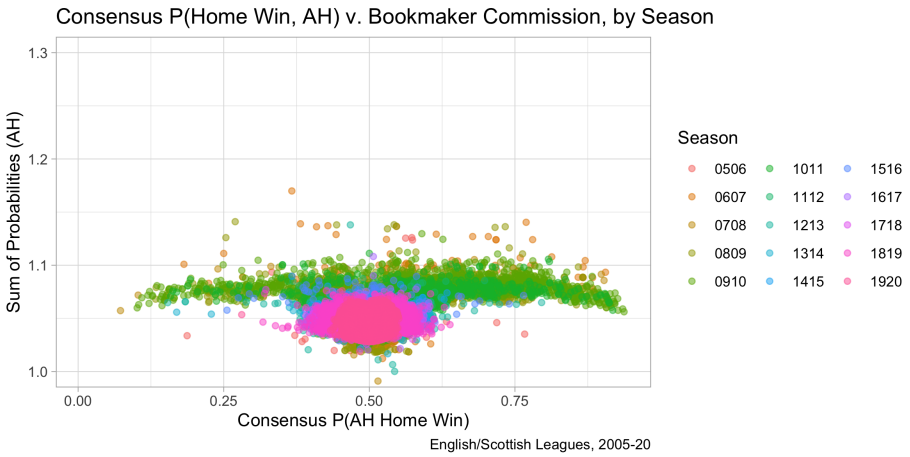


Figure 3.20: The overround in the AH market vs. $\mathbb{P}_{\text{cons}}$(AH Home Win), split by season.

## 3.6 Conclusion

In this section, it has been found that across the English & Scottish football league pyramids, bookmakers are most accurate in the 1X2 market (excluding performance on Draws), followed by the Asian Handicap market, with poor accuracy in the Under/Over 2.5 Goals market, however, the bookmaker accuracy is in fact decreasing in the AH market, and increasing in the UO market. 1X2 odds are, surprisingly, more accurate in Level 3—the lower leagues involving semi-professional sides—than in Level 2—wholly professional lower leagues; performance in Level 1 was greater than all other levels. Finally, in Section 3.5, it has shown the bookmaker commission is highest in the 1X2 market, highest in lower levels, and is reducing over time.

---

In Chapter 4, findings from Chapters 2 and 3 are applied in the creation of a proposed algorithm to place bets in an attempt to use bookmaker accuracy to turn a profit.

# Chapter 4

# A proposed algorithm for placing bets.

In this chapter, findings from Chapters 2 and 3 will be applied to attempt to *beat the bookies* with their own odds, by accepting bookmaker accuracy is high.

## 4.1 The Algorithm

The function in Equation 4.1 is used to choose whether or not to, and how much to, bet in the 1X2 and Asian Handicap markets (measured in arbitrary units) on match $i$, market $m$, denoted as $B_{m,i}$. These markets are chosen as it was found in Chapters 2 and 3 that these show high levels of bookmaker accuracy. Despite the decreasing trend of accuracy in the AH market (Section 3.4) this algorithm is applied on the seasons considered throughout the project: 2005/06 until 2019/20, and as such, this market is chosen. The algorithm is designed to be able to be applied to any market, however, and can easily be adapted to place bets on the Under/Over 2.5 Goals market, or any other offered by bookmakers. The function is based on the consensus probability of the event $p_{m,i}$, where $\mu_m$ is the market mean and $\sigma_m$ is the market standard deviation, both of which are calculated up to and including the match $i$. This is more formally written in Algorithm 1, where $p$ is the set of all consensus probabilities in the market of interest; $n$ the number of matches in the 'information gathering' phase; $N$ the total number of matches in the dataset; $m$ the market chosen (for example, the 1X2 Home Win market).

$$B_{m,i}(p_{m,i}, \mu_m, \sigma_m) = \begin{cases} p_{m,i} < \mu_m + 0.5\sigma_m & B_{m,i} = 0 \\ \mu_m + 0.5\sigma_m \leq p_{m,i} < \mu_m + \sigma_m & B_{m,i} = 1 \\ \mu_m + \sigma_m \leq p_{m,i} < \mu_m + 1.5\sigma_m & B_{m,i} = 2 \\ \mu_m + 1.5\sigma_m \leq p_{m,i} & B_{m,i} = 3 \end{cases} \tag{4.1}$$

More is staked on probabilities further above the mean, as these will have lower odds, and thus, lower winnings, though the chance of the event occurring is higher. Doing it on the number of standard deviations away allows for different variances across the markets (it was found, for example, that the AH market varied less than the 1X2 market; Section 3.1), though the numbers of standard deviations away is an arbitrary choice. This is applied to the entire dataset considered in the project:

**Algorithm 1** A proposed algorithm for placing bets.

1: **function** BET ($p$; $n$; $N$; $m$)
2: $\mu_n \leftarrow$ mean $\mathbb{P}_{\text{cons}}$(Home Win, matches 1:$n$) ▷ 05/06 season data
3: $\sigma_n \leftarrow$ std dev $\mathbb{P}_{\text{cons}}$(Home Win, matches 1:$n$)
4: $B_i \leftarrow 0$ ▷ The placed bets
5: $i \leftarrow n$
6: **while** $i \leq N$ **do**
7: $\mu_i \leftarrow$ mean $\mathbb{P}_{\text{cons}}$(Home Win, matches 1:$i$) ▷ Update $\mu$ and $\sigma$
8: $\sigma_i \leftarrow$ std dev $\mathbb{P}_{\text{cons}}$(Home Win, matches 1:$i$)
9: **if** $p_i > \mu_i + 0.5 \times \sigma_i$ **then**
10: **if** $p_i \leq \mu_i + \sigma_i$ **then**
11: $B_i \leftarrow 1$
12: **else if** $p_i > \mu_i + \sigma_i$ **and** $p_i \leq \mu_i + 1.5 \times \sigma_i$ **then**
13: $B_i \leftarrow 2$
14: **else** ▷ $p_i > \mu_i + 1.5 \times \sigma_i$
15: $B_i \leftarrow 3$
16: **else** ▷ $p_i \leq \mu_i + 0.5 \times \sigma_i$
17: $B_i \leftarrow 0$
18: $i \leftarrow i + 1$ ▷ Iteration step
19: **return** $B_i$ ▷ Repeat for all markets

all markets in the *elite* European leagues[1] and the English & Scottish league pyramids. This is a total of $N = 75086$ matches with $n = 5012$ matches in the 2005/06 season ('information gathering', relating to lines 2–4 of the algorithm).

The algorithm is ran in R using the code below. (N.B., before this is carried out, the data is first read in, using the same method as in Chapters 2 and 3, calling the dataframe `matches`). Statistics for the number of bets, and units staked, is given in Table 4.1.

```
matches$OTHomeBet <- with(matches, 0); matches$OTAwayBet <- with(matches, 0)
matches$AHHomeBet <- with(matches, 0); matches$AHAwayBet <- with(matches, 0)

#Initial Bounds :-
matches0506 <- matches[matches$Season == '0506',]
mu.oth <- mean(matches0506$OT.HProb); sd.oth <- sd(matches0506$OT.HProb)
mu.ota <- mean(matches0506$OT.AProb); sd.ota <- sd(matches0506$OT.AProb)
mu.ahh <- mean(matches0506$AH.HProb); sd.ahh <- sd(matches0506$AH.HProb)
mu.aha <- mean(matches0506$AH.AProb); sd.aha <- sd(matches0506$AH.AProb)
n <- nrow(matches[matches$Season == "0506",]); N <- nrow(matches)
#Placing Bets:-
for (i in n:N){
    #Update the mean and std dev's with our new information
    mu.oth <- mean(matches$OT.HProb[1:i]); sd.oth <- sd(matches$OT.HProb[1:i])
    mu.ota <- mean(matches$OT.AProb[1:i]); sd.ota <- sd(matches$OT.AProb[1:i])
    mu.ahh <- mean(matches$AH.HProb[1:i]); sd.ahh <- sd(matches$AH.HProb[1:i])
    mu.aha <- mean(matches$AH.AProb[1:i]); sd.aha <- sd(matches$AH.AProb[1:i])
    #Do we bet on Home Win (1X2)?
```

[1] Analysis wasn't conducted on the Asian Handicap market in the *elite* dataset, meaning the AH bets placed on those matches will be 'blind' and thus, more fitting to a real-life scenario.

```
    if (matches$OT.HProb[i] > mu.oth + 0.5*sd.oth){
        if (matches$OT.HProb[i] <= mu.oth + sd.oth){matches$OTHomeBet[i] <- 1}
        else if (matches$OT.HProb[i] > mu.oth + sd.oth & matches$OT.HProb[i] <= mu.oth +
            1.5*sd.oth){matches$OTHomeBet[i] <- 2}
        else {matches$OTHomeBet[i] <- 3}}
    else {matches$OTHomeBet[i] <- 0}
    #Do we bet on Away Win (1X2)?
    if (matches$OT.AProb[i] > mu.ota + 0.5*sd.ota){
        if (matches$OT.AProb[i] <= mu.ota + sd.ota){matches$OTAwayBet[i] <- 1}
        else if (matches$OT.AProb[i] > mu.ota + sd.ota & matches$OT.AProb[i] <= mu.ota +
            1.5*sd.ota){matches$OTAwayBet[i] <- 2}
        else {matches$OTAwayBet[i] <- 3}}
    else {matches$OTAwayBet[i] <- 0}
    #Do we bet on Home Win (AH)?
    if (matches$AH.HProb[i] > mu.ahh + 0.5*sd.ahh){
        if (matches$AH.HProb[i] <= mu.ahh + sd.ahh){matches$AHHomeBet[i] <- 1}
        else if (matches$AH.HProb[i] > mu.ahh + sd.ahh & matches$AH.HProb[i] <= mu.ahh +
            1.5*sd.ahh){matches$AHHomeBet[i] <- 2}
        else {matches$AHHomeBet[i] <- 3}}
    else {matches$AHHomeBet[i] <- 0}
    #Do we bet on Away Win (AH)?
    if (matches$AH.AProb[i] > mu.aha + 0.5*sd.aha){
        if (matches$AH.AProb[i] <= mu.aha + sd.aha){matches$AHAwayBet[i] <- 1}
        else if (matches$AH.AProb[i] > mu.aha + sd.aha & matches$AH.AProb[i] <= mu.aha +
            1.5*sd.aha){matches$AHAwayBet[i] <- 2}
        else {matches$AHAwayBet[i] <- 3}}
    else {matches$AHAwayBet[i] <- 0}
}
```

Table 4.1: The total number of bets placed, in units, across each market in the proposed algorithm.

| | *Market* | | | |
|---|---|---|---|---|
| | 1X2 | | AH | |
| *Units Bet* | H Win | A Win | H Win | A Win |
| 0 | 55352 | 55068 | 63348 | 55432 |
| 1 | 8664 | 8880 | 6798 | 14381 |
| 2 | 4951 | 4717 | 2011 | 3847 |
| 3 | 6119 | 6421 | 2929 | 1426 |
| *Total Bets* | 19734 | 20018 | 11738 | 19654 |
| *Stake* | 36923 | 37577 | 19607 | 26353 |

Running the code generates four new columns: `OTHomeBet`, `OTAwayBet`, `AHHomeBet`, and `AHAwayBet` representing $B_{i,\text{OTH}}$, $B_{i,\text{OTA}}$, $B_{i,\text{AHH}}$, $B_{i,\text{AHA}}$ respectively.[2] Each entry is equal to either 0, 1, 2, or 3.

[2]The 1X2 market is represented with `ot` and the Asian Handicap with `ah`, followed by `h` or `a` for the Home or Away Win market.

### 4.1.1 Motivation

Kaunitz, Zhong, and Kreiner (2017) created a high-performing method that made a 3.5% return on 56,435 bets had an accuracy (proportion of bets won) of 44.4%. The paper found mispriced odds by comparing the maximum odds available with the consensus bookmaker odds. If the maximum was sufficiently high (i.e., the consensus probability sufficiently low), a bet (by comparison, worth one unit of the bets placed in this algorithm) was placed. The data used was over a similar timeframe: from 2005 to 2015, but over a much larger set of matches (479,440 matches across 818 leagues). Whilst this method is profitable, individual bettors without the help of methods to deal with large amounts of data (such as data scraping) and methods to find maximum odds and consensus odds will be unable to run this (for example, the `football-data.co.uk` data is uploaded after the matches take place[3] so a bettor would have to collect the information themselves). A bettor would be able to follow the algorithm proposed in this project (a bookmaker, or range of bookmakers, would need to be chosen, and the means/standard deviations can be calculated weekly, monthly, or even by-season). The aim is to create a simpler method that bettors can use based on the findings laid out in this paper.

### 4.1.2 An Alternate Method Excluding Poor Performing Leagues

In addition to finding that the overall bookmaker accuracy was high in the 1X2 Home Win and Away Win, and Asian Handicap markets, two of the *elite* leagues and one of the levels was found to have worse bookmaker performance: the French Ligue Une, German Bundesliga, and the English & Scottish Level 2 (English Leagues One and Two; Scottish Championship). The *alternate* method used will follow the same algorithm, with the exclusion of these leagues. The bets are placed using the code below.

```
matches$OTHomeBet.alt <- with(matches, OTHomeBet)
matches$OTAwayBet.alt <- with(matches, OTAwayBet)
matches$AHHomeBet.alt <- with(matches, AHHomeBet)
matches$AHAwayBet.alt <- with(matches, AHAwayBet)

#Removing bets placed in France, Germany and Level 2, Eng/Sco:-
for (i in 1:N){
    if (matches$Div[i] %in% c("D1", "F1", "E2", "E3", "SC1")){
        matches$OTHomeBet.alt[i] <- 0
        matches$OTAwayBet.alt[i] <- 0
        matches$AHHomeBet.alt[i] <- 0
        matches$AHAwayBet.alt[i] <- 0
    }
}
```

Naturally, the number of bets (and thus, the stake) will be lower in this method; to compare, the accuracy percentage is used, that is, the number of correct bets divided by the total number of bets; and the proportion: the winnings divided by the stake.

[3] Odds are collected, for weekend games, on Friday afternoons, as mentioned in Section 1.3.2, but the final `.csv` files contain match statistics only available after the matches take place.

## 4.2 Computing the Winnings

Once the bets have been placed, the winnings must be found, using the original bookmaker odds offered. Algorithms 2 and 3 are used to calculate the 1X2 and AH winnings respectively, assuming the Winning Probabilities code in Section 3.1 has been ran.

**Algorithm 2** An algorithm for computing the winnings from Algorithm 1 in the 1X2 market.

**function** 1X2WINNINGS ($B_{i,m}$; $O_{i,m}$; $\text{FTR}_i$) ▷ Assuming European odds
  Set all winnings $\omega_{i,m} \leftarrow 0$
  Set market cumulative returns $\Omega_{i,m} \leftarrow 0$
  Initialise count, $i \leftarrow n + 1$ ▷ $n = 5012$, ignore information gathering
  **while** $i \leq N$ **do** ▷ $N = 75086$
    **if** $\text{FTR}_i$ = HomeWin **then**
      $\omega_{i,\text{OTH}} \leftarrow (O_{i,\text{OTH}} - 1) \times B_{i,\text{OTH}}$
      $\omega_{i,\text{OTA}} \leftarrow -B_{i,\text{OTA}}$
    **else if** $\text{FTR}_i$ = AwayWin **then**
      $\omega_{i,\text{OTH}} \leftarrow -B_{i,\text{OTH}}$
      $\omega_{i,\text{OTA}} \leftarrow (O_{i,\text{OTA}} - 1) \times B_{i,\text{OTA}}$
    **else** ▷ I.E., a Draw
      $\omega_{i,\text{OTH}} \leftarrow -B_{i,\text{OTH}}$
      $\omega_{i,\text{OTA}} \leftarrow -B_{i,\text{OTA}}$
    $i \leftarrow i + 1$ ▷ Iteration step
    $\Omega_{i,\text{OTH}} \leftarrow \Omega_{(i-1),\text{OTH}} + \omega_{i,\text{OTH}}$ ▷ Update cumulative returns
    $\Omega_{i,\text{OTA}} \leftarrow \Omega_{(i-1),\text{OTA}} + \omega_{i,\text{OTA}}$
  **return** $\Omega_{\text{OTH}}$, $\Omega_{\text{OTA}}$

**Algorithm 3** An algorithm for computing the winnings from Algorithm 1 in the AH market.

**function** AHWINNINGS ($B_{i,m}$; $O_{i,m}$; $\text{AHRes}_i$) ▷ Assuming European odds
  Set all winnings $\omega_{i,m} = 0$
  Set market cumulative returns $\Omega_{i,m} \leftarrow 0$
  Initialise count, $i \leftarrow n + 1$ ▷ $n$ and $N$ as before
  **while** $i \leq N$ **do**
    **if** $\text{AHRes}_i = \text{FullHomeWin}$ **then**
      $\omega_{i,\text{AHH}} \leftarrow (O_{i,\text{AHH}} - 1) \times B_{i,\text{AHH}}$
      $\omega_{i,\text{AHA}} \leftarrow -B_{i,\text{AHA}}$
    **else if** $\text{AHRes}_i = \text{HalfHomeWin}$ **then**
      $\omega_{i,\text{AHH}} \leftarrow (O_{i,\text{AHH}} - 1) \times 0.5 \times B_{i,\text{AHH}} - (0.5 \times B_{i,\text{AHH}})$
      $\omega_{i,\text{AHA}} \leftarrow -B_{i,\text{AHA}}$
    **else if** $\text{AHRes}_i = \text{FullAwayWin}$ **then**
      $\omega_{i,\text{AHH}} \leftarrow -B_{i,\text{AHH}}$
      $\omega_{i,\text{AHA}} \leftarrow (O_{i,\text{AHA}} - 1) \times B_{i,\text{AHA}}$
    **else if** $\text{AHRes}_i = \text{HalfAwayWin}$ **then**
      $\omega_{i,\text{AHH}} \leftarrow -B_{i,\text{AHH}}$
      $\omega_{i,\text{AHA}} \leftarrow (O_{i,\text{AHA}} - 1) \times 0.5 \times B_{i,\text{AHA}} - (0.5 \times B_{i,\text{AHA}})$
    **else** ▷ I.E., a Voided Bet
      $\omega_{i,\text{AHH}} \leftarrow 0$
      $\omega_{i,\text{AHA}} \leftarrow 0$
    $i \leftarrow i + 1$ ▷ Iteration step
    $\Omega_{i,\text{AHH}} \leftarrow \Omega_{(i-1),\text{AHH}} + \omega_{i,\text{AHH}}$ ▷ Update cumulative returns
    $\Omega_{i,\text{AHA}} \leftarrow \Omega_{(i-1),\text{AHA}} + \omega_{i,\text{AHA}}$
  **return** $\Omega_{\text{AHH}},\ \Omega_{\text{AHA}}$

(N.B., One may notice that one is subtracted from the odds offered, and wonder why. As mentioned in Section 1.1.2, the decimal (European) odds are defined with the stake included, and would therefore inflate the winnings reported. Strict winnings can be found using the fractional (British) odds, computed by subtracting one from the decimal odds. If fractional odds were inputted to the algorithm, these steps should be ignored.)

## 4.3 Results

In addition to running the algorithms in R, each match is assigned with an 'index' by taking matches $n+1$ to $N$ and numbering them from 1 to $N-n$. The index is used to create a line plot of the cumulative winnings. This is shown in Figure 4.1, with a zoomed-in plot of the first 100 matches shown in Figure 4.2.

The performance of the first 100 matches is not too bad, with the overall profit line hovering just below the $y = 0$ line. The total winnings, however, show a rather conclusive loss. 'Money' is lost in all four markets, with the biggest loss ($\approx 2,000$ units) in the Asian Handicap Away Win market. Full figures are in Table 4.2.

Table 4.2: The proposed betting algorithm's winnings and accuracy.

| *Market* | *Bets* | *Stake* | *Winnings* | *Proportion* | *Accuracy (%)* |
|---|---|---|---|---|---|
| 1x2 H | 19734 | 36923 | –1103.380 | –0.030 | 63.363 |
| 1x2 A | 20018 | 37577 | –1184.610 | –0.032 | 47.472 |
| AH H | 11738 | 19607 | –956.325 | –0.049 | 45.604 |
| AH A | 19654 | 26353 | –1988.525 | –0.075 | 43.625 |
| Overall | 71144 | 120460 | –5232.84 | –0.043 | 50.509 |

The table shows an overall 4.3% loss of the stake, with a loss of around 3% for the two 1X2 markets, and a higher loss of 5% and 7.5% for the two AH markets. Despite these losses, the algorithm has an exceptionally high accuracy: over 50% of all bets placed were winners, with an accuracy of 63% for the 1X2 Home Win market. Compared against Kaunitz, Zhong, and Kreiner (2017)'s method discussed in Section 4.1.1, only the AH Away Win market has a lower accuracy, but the proposed algorithm's returns are 7.8% lower.

The same values for the *alternate* method defined in Section 4.1.2 are found and presented in Table 4.3.

Table 4.3: The *alternate* betting model winnings and accuracy, with comparisons against values in Table 4.2.

| *Market* | *Bets* | *Stake* | *Winnings* | *Proportion* | *Accuracy (%)* |
|---|---|---|---|---|---|
| 1x2 H | 12812 | 24974 | –536.70 | –0.021 (+0.008) | 64.924 (+1.561) |
| 1x2 A | 12632 | 24808 | –698.10 | –0.028 (+0.003) | 48.789 (+1.317) |
| AH H | 7444 | 12713 | –566.08 | –0.045 (+0.004) | 45.621 (+0.017) |
| AH A | 11852 | 16094 | –1268.31 | –0.079 (–0.003) | 43.495 (–0.130) |
| Overall | 44740 | 78589 | –3069.19 | –0.039 (+0.004) | 51.480 (+0.971) |

Whilst all markets but the AH Away Win market showed an increased proportion of winnings and accuracy, these were not by much, with the former under 1% in all cases. This shows that using the same method, but ignoring the worst performing leagues, does not increase the profitability of the algorithm proposed.

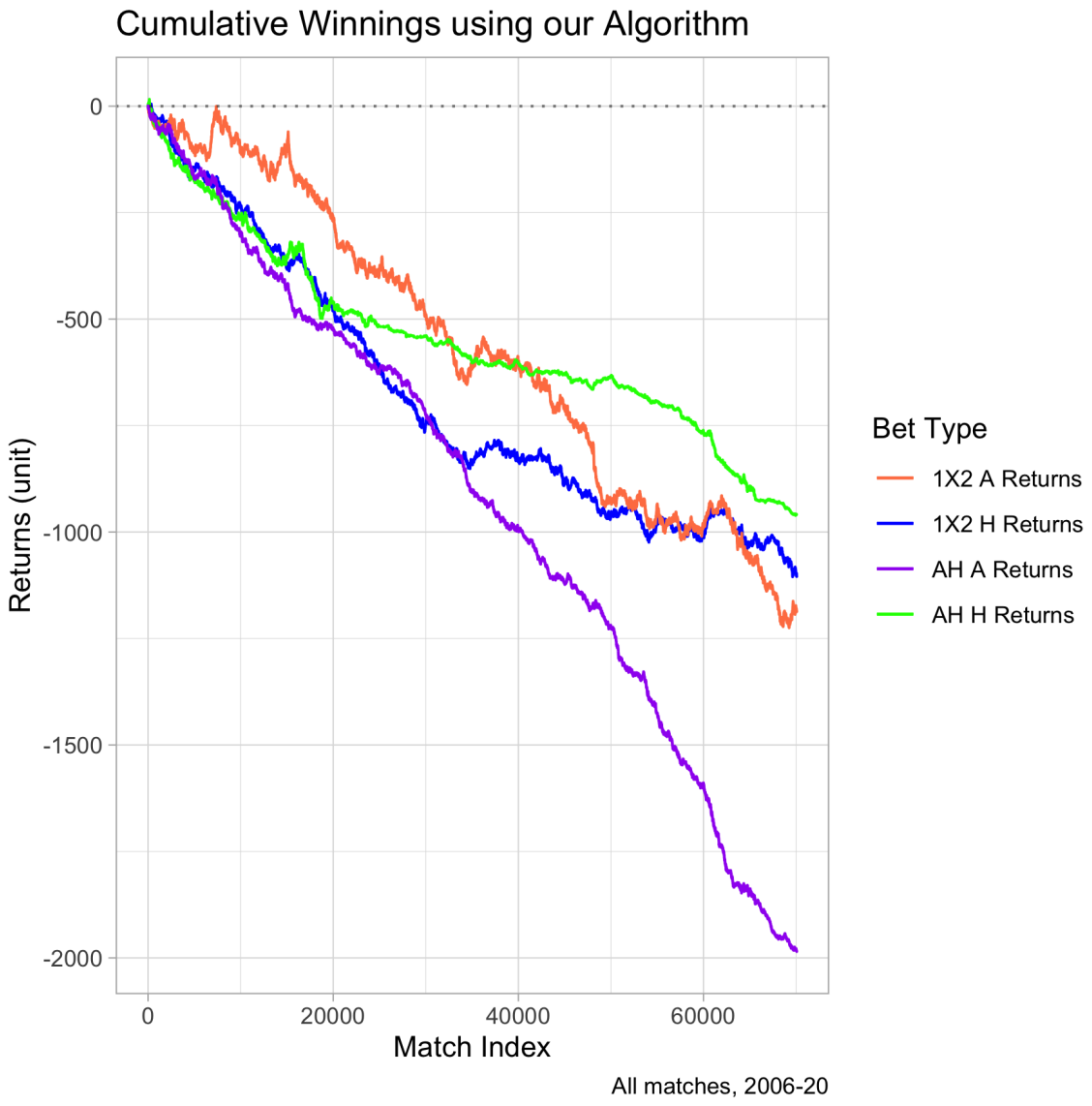


Figure 4.1: A plot of the cumulative winnings of the proposed betting algorithm, split by market.

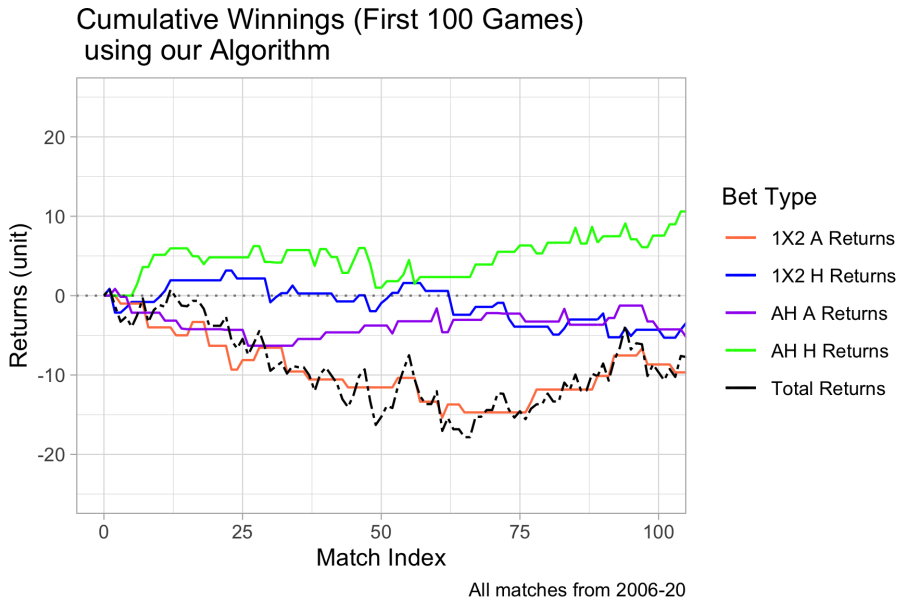


Figure 4.2: A zoomed in view of Figure 4.1, considering the first 100 matches.

## 4.4 Comparison Against a Random Bet Strategy

To further assess the algorithm, it is compared against a random bet strategy (RBS) which will place bets with the same distribution as the proposed algorithm, but on random matches. The probabilities of a 1 unit, 2 unit, or 3 unit stake are found by dividing the number of such bets (Table 4.1) by $N - n = 70074$. For example,

$$\begin{aligned} \mathbb{P}(\text{Stake} = 2 \text{ in 1X2 Home Win market}) &= 4951/70074 \\ &= 0.0707 \end{aligned}$$

Using R, the four probabilities (Unit Bet = 0, 1, 2, or 3) can be computed using the code below, where `BET` is changed to the desired market.

```
p1 <- nrow(matches[matches$OTHomeBet == 1,]) / nrow(matches[(n+1):N,])
p2 <- nrow(matches[matches$OTHomeBet == 2,]) / nrow(matches[(n+1):N,])
p3 <- nrow(matches[matches$OTHomeBet == 3,]) / nrow(matches[(n+1):N,])
p0 <- 1 - (p1 + p2 + p3)
```

Matches are assigned their bets using the `e1071` package, which allows for the creation of a discrete probability distribution with defined probabilities (Meyer et al., 2020). The code below is used to do this, where the first line resets the bets (as this is done in a `for` loop). Once this is completed, the winning bets and returns are found as before.

```
matches$rand.Bet <- with(matches, 0)
matches$rand.Bet <- with(matches, rand.Bet.OTH + rdiscrete(n = nrow(matches), values =
    0:3, probs=c(p0, p1, p2, p3)))
```

The RBS is used to create accuracy tables as in Table 4.2. After being ran ten times, to ensure results are representative, averages are taken and given in Table 4.4. The values from ten individual runs are given in Appendix E.

Table 4.4: Average values from 10 runs of the random bet strategy winnings and accuracy, with comparisons against values in Table 4.2.

| *Market* | *Bets* | *Stake* | *Winnings* | *Proportion* | *Accuracy (%)* |
|---|---|---|---|---|---|
| 1x2 H | 19652.1 | 39431.8 | –2288.91 | –0.058 (–0.028) | 44.231 (–19.131) |
| 1x2 A | 20012.6 | 40284.6 | –3182.22 | –0.079 (–0.047) | 30.038 (–17.434) |
| AH H | 11723.3 | 20986.9 | –3734.98 | –0.178 (–0.129) | 41.483 (–4.121) |
| AH A | 19700.6 | 28330.5 | –5018.98 | –0.177 (–0.102) | 40.319 (–3.305) |
| Overall | 71088.6 | 129033.8 | –14225.09 | –0.110 (–0.067) | 38.699 (–11.810) |

A line plot similar to that in Figure 4.1 is produced with the RBS ran thirty times, split into each market to allow for in-depth comparisons, used in addition to Table 4.4. These are shown in Figures 4.3 to 4.6.

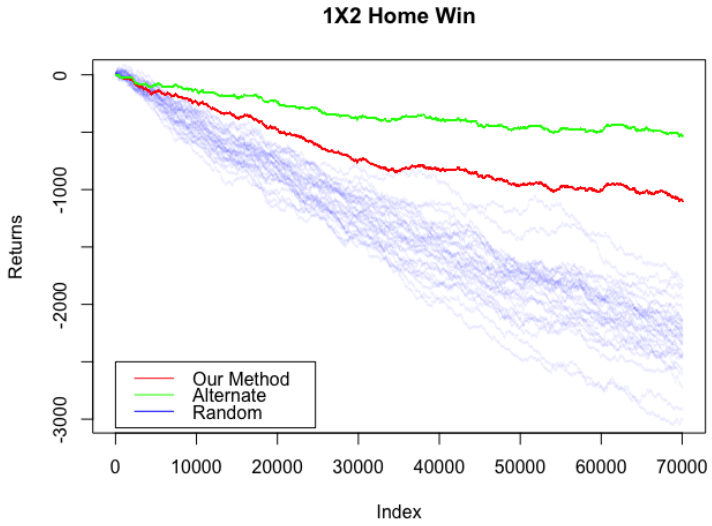


Figure 4.3: A line graph of the winnings of the random bet strategy in the 1X2 Home Win market, compared to the proposed algorithm and the *alternate* method.

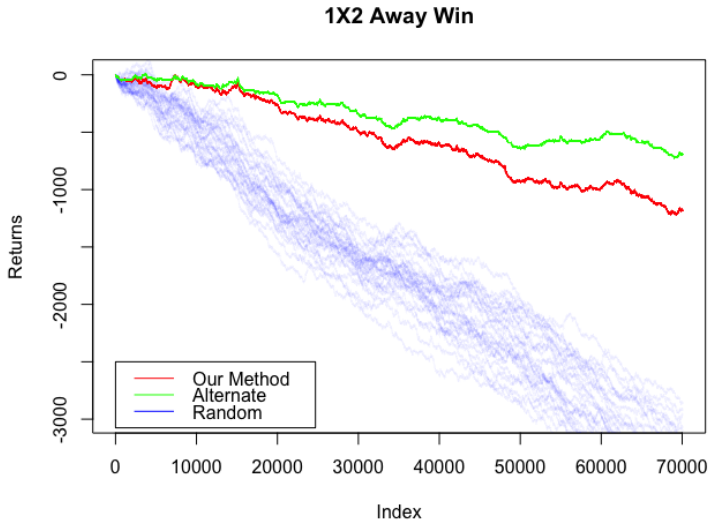


Figure 4.4: A line graph of the winnings of the random bet strategy in the 1X2 Away Win market, compared to the proposed algorithm and the *alternate* method.

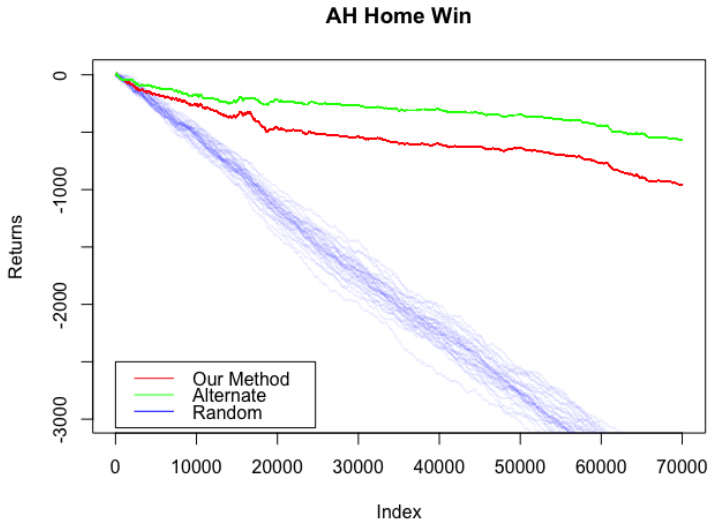


Figure 4.5: A line graph of the winnings of the random bet strategy in the AH Home Win market, compared to the proposed algorithm and the *alternate* method.

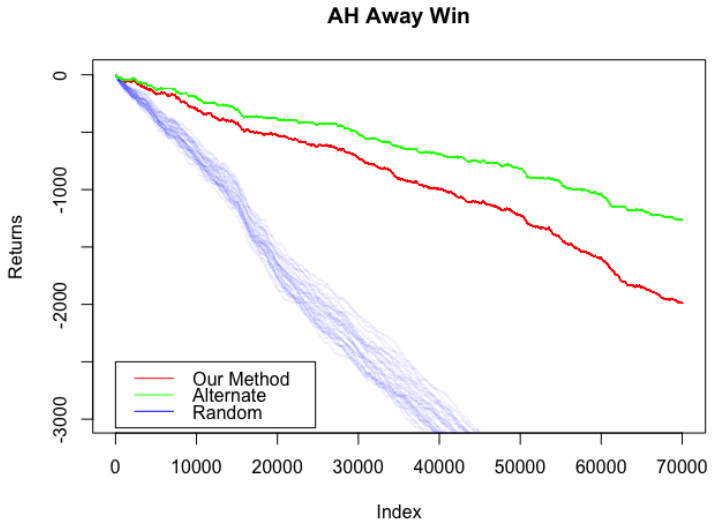


Figure 4.6: A line graph of the winnings of the random bet strategy in the AH Away Win market, compared to the proposed algorithm and the *alternate* method.

The figures, and Table 4.4, show a clear trend. Whilst the proposed algorithm makes a loss, it is significantly better than the RBS, with the most significant improvements coming in the AH market. Figures 4.5 and 4.6 display this, with a immediate and instant loss with the RBS. On average, the RBS lost 17.8% and 17.7% of the stake in the Asian Handicap markets, compared to 4.9% and 7.5% for the proposed algorithm in the corresponding markets: a difference of 12.9% and 10.2%. The accuracy, however, had less than 4% difference in each market.

Interestingly, the line for the RBS returns in the AH market had far less variation than the 1X2 market, which could suggest that, unless the bettor has a set strategy and/or insider knowledge, a loss of around 17.7% is inevitable.

In the 1X2 market, despite vast decreases in accuracy (the RBS for the Home Win market was 19% less accurate; the Away Win market was 17% less accurate), the final winnings were not as poor as in the AH markets, with 2.8% and 4.7% larger losses. In addition, from the figures, the RBS has a larger distribution in the 1X2 markets, with some runs performing similarly to the proposed algorithm for a long time: in Figure 4.3, for example, two runs with similar results to the algorithm are observed until around Match Index 50000. The proposed algorithm always finishes more profitable than the RBS, however.

## 4.5 Conclusion

The proposed algorithm outperforms the random bet strategy. In order to improve the results, with an aim to create a profit, however, more advanced methods, such as utilising maximum odds on offer for a match, vast amounts of data, or insider knowledge are required. It has been shown that a high level of accuracy doesn't equate to winnings, or that a vastly lower accuracy results in much lower profit. In addition, it was shown that the Asian Handicap market was harder to make winnings from than the 1X2 market, without either a being knowledgeable bettor, and/or utilising a more advanced model.

It is, therefore, unsurprising that the bookmakers are so profitable: Entain—the owners of Coral and Ladbrokes—reported a £175 million profit in the year ending in 2020; Flutter—the owner of PaddyPower, Sportsbet, Betfair, and SkyBet—reported a profit of £136 million in 2019; and bet365 reported a profit of £560 million in 2019 (Davies, 2021; Flutter Entertainment, plc., 2020; bet365 Group Limited, 2019). In addition, bookmakers are known to discriminate against successful bettors, often refusing bets to be placed, or even refusing to pay out large sums, thus making it even hard for bettors to make profits (Symonds, 2020; Cave, 2015; Osborne, 2015).

---

In Chapter 5, the findings made throughout the project are given together, as well as a personal reflection into what I have learnt throughout the course of the project.

# Chapter 5

# Conclusion

In this final chapter, the findings from this project will be discussed as well as areas for future research, and, from a self-reflecting view, challenges I have been faced with throughout the duration of the project, and how I overcame them.

## 5.1 Findings, Strengths, and Limitations

The aim of this project was to use a range of measures and methods to quantify bookmaker accuracy in football, across two different league groups (an *elite* group of six top-tier European leagues and between different levels of the English & Scottish pyramids), and across three popular markets (the 1X2, Under/Over 2.5 Goals (UO), and Asian Handicap (AH)). This was paired with investigations into the effect of competitive balance on accuracy and into the overround—a measure of bookmaker commission—set by bookmakers.

### The 1X2 Market

It has been shown that the bookmaker accuracy is high in the 1X2 Home Win and Away Win markets, based on visual analyses, and five measures of statistical accuracy: $R^2$, RMSE, $P_1$, $P_2$, and the Slope of the linear model created to find $R^2$ and RMSE. Bookmakers, however, struggle with the prediction of Draws, and opt to use a 'safe' method of setting relatively constant odds (the consensus probabilities have a low standard deviation), set to reflect the actual probability of a Draw occurring (approximately between 0.25 and 0.27, changing over time and across leagues). These findings are as expected. What is, perhaps, more unexpected is in the English & Scottish pyramid analysis, bookmaker accuracy in all four measures was worse in the Level 2 group of leagues (English Leagues One and Two; Scottish Championship) than in the Level 3 group (English National League; Scottish Leagues One and Two).

### The Under/Over 2.5 Goals Market

The UO market was shown to exhibit poor levels of accuracy from bookmakers, with an $R^2$ of 38.7% and RMSE of 0.1478. It was shown to be improving over time, with four $R^2$ values greater than 75% in the last five seasons considered. $P_1$ and $P_2$ steadily improved too, with the former rising from 0.5020 to 0.5039, and the latter decreasing from 0.4960 to 0.4923. Between levels, in the $R^2$ and RMSE variables, Level 1 and 2 appear to have similar performance: $R^2_{L1} = 64\%$, $R^2_{L2} = 53\%$ and

$\text{RMSE}_{L1} = 0.1047$, $\text{RMSE}_{L2} = 0.1241$. Level 3 had much worse performance by these two measures, but appeared to perform better than Level 2 based on the $P$-values.

### The Asian Handicap Market

It was shown via visual analysis (Figure 3.7) that the handicap is well placed and is improving over time (the variance in odds offered is reducing, and approaching 0.5, ideal AH odds; Figure 3.20). However, the market performance appears to be decreasing over time, with $R^2$ decreasing (in 2005/06, $R^2 = 81\%$ to, in 2019/20, $R^2 = 53\%$), RMSE increasing (0.1568 to 0.1605). The $P$-values appear to be worsening over time, too, with $P_1$ dropping from 0.5026 to 0.5006, and $P_2$ rising from 0.4948 to 0.4988. These values, however, still indicate slightly better performance in this market than in the UO market: the $R^2$ and RMSE values may be unreliable due to the lower variation in consensus probabilities making a linear model unsuitable. Between levels, bookmakers performed best in Level 1, across all four variables, followed by Levels 2 and 3, with a decrease in $R^2$ and $P_1$, and an increase in RMSE and $P_2$.

### The Effect of Competitive Balance

Competitive balance affects bookmaker accuracy. Three measures were used to quantify competitive balance; it was shown in leagues where balance is high (such as the French Ligue Une and German Bundesliga), bookmakers, whilst performing highly, performed relatively worse than in leagues where balance is low (Portuguese Primiera Liga).

### Bookmaker Overround

Bookmakers set a higher overround at lower levels; this is more evident in the 1X2 and UO markets. Over time, the overround is reducing in these two markets, whilst remaining stable in the AH market. It was found the AH handicap choice is improving, as the variation of consensus probabilities (and hence, odds offered) is reducing, and concentrating around 0.5.

### Findings from the Proposed Betting Algorithm

Whilst it did not perform as well as hoped, and ultimately made a loss in all four markets, a connection was made between the percentage of bets won, and the actual winnings of the bets. The algorithm won more bets than the successful method proposed by Kaunitz, Zhong, and Kreiner (2017) yet lost a much greater percentage. To win money whilst betting, bettors without 'insider' knowledge therefore likely require large amounts of data (Godin et al., 2014) or advanced mathematical tools (Dixon and Coles, 1997; Owen, 2009; Karlis and Ntzoufras, 2009; Constantinou, 2020).

### Strengths and Limitations of this Study

This project established a possible link between competitive balance and accuracy, and reviewed a wide range of leagues, across multiple markets, using various measures.

Naturally, there are limitations to this piece of work. Firstly, the AH and UO markets could be assessed for the *elite* group.[1] In addition, the data was from a limited range of seasons, 2005/06 to 2019/20, and with only a select group of leagues; for the investigation into bookmaker accuracy across

[1] This was not done to create a 'blind' environment for the algorithm in Chapter 4.

levels, perhaps finding more data would have assisted the results. The $R^2$ and RMSE values used throughout are based on the creation of linear models: this may not have been the most appropriate choice of model, or the most appropriate choice of statistics. RMSE, for instance, computes the distance of a datapoint to the linear model, rather than the 'perfect accuracy' $x = y$ line. The slope of the linear model is considered instead, though if a linear model is inappropriate, the slope is also inapplicable.

For the investigation into competitive balance, Goossens (2005) uses data from the 1963/64 season until the 2004/05 season: the findings would be improved if the data coincided with the same seasons. In addition, the Top $k$ Teams statistic (named the kappa $\kappa$ value) uses an arbitrary choice of $k = 3$ which may not accurately reflect balance over time: the English Premier League, for example, had a period of having a 'big four' teams in the 2000s (Arsenal, Chelsea, Liverpool, Manchester United), before an ongoing period of a 'top six' (with the additions of Manchester City and Tottenham Hotspur) (Kelly, 2021).

Of course, this project has also left stones unturned and more questions to be asked, such as:

- Does there exist a relationship between the style of play of a football league (whether the league is generally more offensive or defensive, or individual teams) and bookmaker accuracy?
- Has the COVID-19 pandemic affected bookmaker accuracy, or football itself (with changes to home advantage, for example)?
- Could it be possible for bookmakers to accuracy predict Draws?
- Is the competitive balance of the English & Scottish lower professional leagues higher than that of the top-tiers, and semi-professional football? (This could explain why bookmaker accuracy is lower in the Level 2 leagues in Chapter 3.)

## 5.2 Challenges

This project has tested my understanding of probabilities, odds, and various statistical techniques and tools. I used R to conduct my analysis, and my knowledge of which—despite having used it in several modules throughout my degree—has grown enormously over the course of the project, for example, dealing with programming problems and learning how to solve the plethora of issues that can occur, from simple syntax errors to issues with data types. At the start of the project, I was downloading every `.csv` file from `football-data.co.uk` to use within my coding, which was very time consuming. Efficient methods, such as using `for` loops regularly within the code helped massively. In addition, the use of the `ggplot2` package has allowed me to create both a wider range of, and more aesthetically pleasing, plots: prior to the project, I had never used it before. Perhaps the most challenging code I used is the code to run the random bet strategy (Section 4.4), where I defined a new distribution (using the `discrete` tool from the `e1071` package) and ran the code numerous times: first to create the tables in Appendix E, second to create the plots in Figures 4.3 to 4.6.

Formal critical analysis of literature has been something I have never done prior to the project, and something I initially struggled with. I found, after several constructive meetings, good techniques both for creating the Literature Review (Section 1.2) and for using findings within the project.

Project management has been crucial, too. Whilst the overall time frame is long, I found setting myself small goals and targets weekly or biweekly a great way of keeping on track, and continually making progress. Unlike other courseworks throughout university, there has been no set mark scheme/list of questions to follow, rather I have had to set the questions and aims myself.

Finally, this dissertation has tested both my LaTeX skills, and communication, trying to ensure I present my findings clearly, concisely, and without unnecessary repetition. I hope I have managed to do this!

# Bibliography

Adèr, Herman J. (2008). "Phases and initial steps in data analysis". In: *Advising on Research Methods: A consultant's companion*. Chap. 14, pp. 333–357.

Ajadi, Theo et al. (2020). *Home truths: Annual Review of Football Finance 2020*. URL: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-business-group/deloitte-uk-annual-review-of-football-finance-2020.pdf (visited on 03/25/2021).

Alto, Valentina (2019). *PCA: Eigenvectors and Eigenvalues*. URL: https://towardsdatascience.com/pca-eigenvectors-and-eigenvalues-1f968bc6777a (visited on 03/24/2021).

American Psychiatric Association (2018). "DSM5 Diagnostic Criteria: Gambling Disorder". In: *Diagnostic and Statistical Manual of Mental Disorders (5th Edition)*.

Angelini, Giovanni and Luca De Angelis (2019). "Efficiency of online football betting markets". In: *International Journal of Forecasting* 35.2, pp. 712–721.

Auguie, Baptiste (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. URL: https://CRAN.R-project.org/package=gridExtra.

Barrabi, Thomas (2020). *How does NFL's salary cap work?* URL: https://www.foxbusiness.com/sports/nfl-salary-cap-rules-explained (visited on 03/21/2021).

bet365 (n.d.). *Help: Soccer*. URL: https://help.bet365.com/product-help/sports/Rules/Soccer (visited on 01/12/2021).

bet365 Group Limited (2019). *Report and Financial Statements*. URL: https://find-and-update.company-information.service.gov.uk/company/04241161/filing-history/MzI1MjIzNDU1MGFkaXF6a2N4/document?format=pdf (visited on 04/13/2021).

Boscá, José E et al. (2009). "Increasing offensive or defensive efficiency? An analysis of Italian and Spanish football". In: *Omega* 37.1, pp. 63–78.

Buchdahl, Joseph (n.d.[a]). *Notes for Football Data*. URL: http://www.football-data.co.uk/notes.txt (visited on 10/02/2020).

— (n.d.[b]). *Pinnacle Betting Expert and Author: Joseph Buchdahl*. Pinnacle. URL: https://www.pinnacle.com/en/betting-resources/author/Joseph-Buchdahl (visited on 01/08/2021).

— (n.d.[c]). *What is Football-Data?* Football Data. URL: http://football-data.co.uk/#intro (visited on 01/08/2021).

Cain, Michael et al. (2000). "The favourite-longshot bias and market efficiency in UK football betting". In: *Scottish Journal of Political Economy* 47.1, pp. 25–36.

Cambridge Dictionary, ed. (n.d.). *COMPETITIVE BALANCE — definition in the Cambridge English Dictionary*. URL: https://dictionary.cambridge.org/us/dictionary/english/competitive-balance (visited on 01/12/2021).

Cave, Rob (2015). *Banned from the bookies*. URL: https://www.bbc.co.uk/news/business-34550617 (visited on 04/13/2021).

Chen, James (2020). *Line Of Best Fit*. URL: https://www.investopedia.com/terms/l/line-of-best-fit.asp (visited on 04/06/2021).

Clapham, Christopher and James Nicholson (2014). *Oxford Concise Dictionary of Mathematics. 5th Edition*. Oxford University Press.

Conn, David (2017). *Premier League finances: the full club-by-club breakdown and verdict*. URL: https://www.theguardian.com/football/2017/jun/01/premier-league-finances-club-by-club (visited on 01/12/2021).

Constantinou, Anthony (2020). *Asian Handicap football betting with Rating-based Hybrid Bayesian Networks*. arXiv: 2003.09384 [stat.AP].

Cortis, Dominic (2015). "Expected values and variances in bookmaker payouts: A theoretical approach towards setting limits on odds". In: *The Journal of Prediction Markets* 9.1, pp. 1–14.

Cronin, Benjamin (2019). *American odds vs. Decimal odds*. Pinnacle. URL: https://www.pinnacle.com/en/betting-articles/educational/odds-formats-available-at-pinnacle-sports/ZWSJD9PPX69V3YXZ (visited on 01/08/2021).

Davey, Jacob (2020). *Why It's Time We Ban Betting Sponsors in English Football*. URL: https://versus.uk.com/2020/07/time-ban-betting-sponsors-english-football/ (visited on 02/15/2021).

Davies, Rob (2021). *UK betting companies report profits in 2020 as US market opens up*. URL: https://www.theguardian.com/society/2021/mar/04/uk-betting-companies-report-profits-in-2020-as-us-market-opens-up (visited on 04/13/2021).

DeepAI (n.d.). *Kernel Density Estimation Definition*. URL: https://deepai.org/machine-learning-glossary-and-terms/kernel-density-estimation (visited on 04/06/2021).

Dixon, Mark J and Stuart G Coles (1997). "Modelling association football scores and inefficiencies in the football betting market". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46.2, pp. 265–280.

Draper, Norman R. and Harry Smith (1998). *Applied Regression Analysis. 3rd Edition*. John Wiley & Sons, New York.

English Football League (n.d.). *EFL Regulations, Section 3 - The League*. URL: https://www.efl.com/-more/governance/efl-rules--regulations/section-3---the-league/ (visited on 03/25/2021).

Fabrigar, Leandre R et al. (1999). "Evaluating the use of exploratory factor analysis in psychological research." In: *Psychological methods* 4.3, p. 272.

FIFA.com (2001). *FIFA Survey: approximately 250 million footballers worldwide*. URL: https://www.fifa.com/who-we-are/news/fifa-survey-approximately-250-million-footballers-worldwide-88048 (visited on 01/12/2021).

Flutter Entertainment, plc. (2020). *2019 Annual Report and Accounts*. URL: https://www.flutter.com/sites/paddy-power-betfair/files/Annual%20reports/2019-annual-report-28-02-20.pdf (visited on 04/13/2021).

footballxg.com (n.d.). *What are Expected Goals (xG)?* URL: https://footballxg.com/what{_}are{_}expected{_}goals/ (visited on 04/24/2021).

Fox, John and Sanford Weisberg (2019). *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Giulianotti, Richard (2012). "Football". In: *The Wiley-Blackwell Encyclopedia of Globalization*. American Cancer Society.

Gladwell, Ben (2015). *Catania president arrested in relation to match-fixing investigation*. URL: https://www.espn.com/soccer/italian-serie-b/story/2502122/catania-president-arrested-in-relation-to-match-fixing (visited on 04/24/2021).

Goal (2019). *What is the MLS salary cap & how much are U.S. soccer players paid?* URL: https://www.goal.com/en-us/news/mls-salary-cap-how-much-us-soccer-players-paid/q015j4su3gb31bha41zto4fkb (visited on 03/21/2021).

Godin, Fréderic et al. (2014). "Beating the bookmakers: leveraging statistics and Twitter microposts for predicting soccer results". In: *Workshop on Large-Scale Sports Analytics (KDD 2014), New York, USA*.

Goossens, Kelly (2005). *Competitive Balance in European Football: Comparison by adapting measures: National Measure of Seasonal Imbalance and Top3*. University of Antwerp, Research Administration.

Grinstead, Charles Miller and James Laurie Snell (2012). *Introduction to probability*. American Mathematical Soc.

Hafez, Shamoon (2019). *Calciopoli: The scandal that rocked Italy and left Juventus in Serie B*. URL: https://www.bbc.co.uk/sport/football/49910626 (visited on 01/07/2021).

Hoaglin, David C. (1977). "Mathematical Software and Exploratory Data Analysis". In: *Mathematical Software*. Ed. by John R. Rice. Academic Press, pp. 139–159. ISBN: 978-0-12-587260-7. DOI: https://doi.org/10.1016/B978-0-12-587260-7.50010-8.

Hoaglin, David C. and Roy E. Welsch (1978). "The Hat Matrix in Regression and ANOVA". In: *The American Statistician* 32.1, pp. 17–22.

Hyndman, Rob and Anne Koehler (2006). "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22, pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

Jolliffe, Ian T (1972). "Discarding variables in a principal component analysis. I: Artificial data". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 21.2, pp. 160–173.

Kaiser, Henry F (1974). "An index of factorial simplicity". In: *Psychometrika* 39.1, pp. 31–36.

Karlis, Dimitris and Ioannis Ntzoufras (2009). "Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference". In: *IMA Journal of Management Mathematics* 20.2, pp. 133–145.

Kaunitz, Lisandro, Shenjun Zhong, and Javier Kreiner (2017). "Beating the bookies with their own numbers - and how the online sports betting market is rigged". In: *arXiv preprint arXiv:1710.02824*.

Kelly, Ryan (2021). *Who are the Premier League 'big six'? Top English clubs & nickname explained.* URL: https://www.goal.com/en-gb/news/who-are-premier-league-big-six-top-english-clubs-nickname/130iokmi8t8dt1k3kudou73s1k (visited on 04/15/2021).

Khazaal, Yasser et al. (2012). "Effects of expertise on football betting". In: *Substance abuse treatment, prevention, and policy* 7.1, pp. 1–6.

Knuth, Kevin H. (2006). "Optimal data-based binning for histograms". In: *arXiv preprint physics/0605197*.

Kuypers, Tim (2000). "Information and efficiency: an empirical study of a fixed odds betting market". In: *Applied Economics* 32.11, pp. 1353–1363.

Lange, David (2020). *Teams of the German Bundesliga ranked by market (transfer) value of players in 2020.* URL: https://www.statista.com/statistics/283033/market-value-teams-german-football-bundesliga/#statisticContainer (visited on 01/12/2021).

— (2021). *Average player salary in the EPL 2019/20, by team.* URL: https://www.statista.com/statistics/675303/average-epl-salary-by-team/ (visited on 04/24/2021).

Levitt, Steven D (2004). "Why are gambling markets organised so differently from financial markets?" In: *The Economic Journal* 114.495, pp. 223–246.

Limited, The Football Association Premier League (2020). *Premier League Football Scores, Results & Season Archives.* URL: https://www.premierleague.com/results?co=1{\&}se=274{\&}cl=1 (visited on 04/24/2021).

Lorenz, M. O. (1905). "Methods of Measuring the Concentration of Wealth". In: *Publications of the American Statistical Association* 9.70, pp. 209–219.

Ma, Shuangge and Ying Dai (2011). "Principal component analysis based methods in bioinformatics studies". In: *Briefings in bioinformatics* 12.6, pp. 714–722.

MatterOfStats (n.d.). *What is Vig and Overround.* URL: http://www.matterofstats.com/what-is-vig-and-overround/ (visited on 04/06/2021).

Mendenhall, William, Robert J. Beaver, and Barbara M. Beaver (2013). *Introduction to Probability and Statistics (14th Edition).* Brooks/Cole.

Meyer, David et al. (2020). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* R package version 1.7-4. URL: https://CRAN.R-project.org/package=e1071.

Morris, Julie A and Martin J Gardner (1988). "Statistics in medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates". In: *British medical journal (Clinical research ed.)* 296.6632, p. 1313.

NBA.com (2020). *NBA Salary Cap set at $109.14 million for 2019-20.* URL: https://www.nba.com/news/nba-salary-cap-2019-20-season-set-10914-million (visited on 03/21/2021).

Osborne, Samuel (2015). *Bookmakers 'refusing to take bets from successful gamblers'.* URL: https://www.independent.co.uk/news/uk/home-news/bookmakers-refusing-take-bets-successful-gamblers-a6698756.html (visited on 04/13/2021).

Owen, Alun (2009). "Dynamic bayesian forecasting models of football match outcomes". In: *2nd International Conference on Mathematics in Sport (IMA Sport 2009)* (Groningen, The Netherlands). Ed. by The Institute of Mathematics and its Applications (IMA).

Pearson, Karl (1901). "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572.

Plotly.com (n.d.). *geom count in ggplot2*. URL: `https://plotly.com/ggplot2/geom{\_}count/` (visited on 03/28/2021).

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Ramchandani, Girish (2012). "Competitiveness of the English Premier League (1992-2010) and ten European football leagues (2010)". In: *International Journal of Performance Analysis in Sport* 12.2, pp. 346–360.

Rathke, Alex (2017). "An examination of expected goals and shot efficiency in soccer". In: *Journal of Human Sport and Exercise* 12.2, pp. 514–529.

RDocumentation (n.d.). *cut: Convert Numeric to Factor*. URL: `https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cut` (visited on 03/21/2021).

Rosen, Dan (2020). *NHL salary cap to remain same next season*. URL: `https://www.nhl.com/news/nhl-salary-cap-to-remain-at-815-million/c-317372082` (visited on 03/21/2021).

Rottenberg, Simon (1956). "The Baseball Players' Labor Market". In: *Journal of Political Economy* 64.3, pp. 242–258. DOI: `10.1086/257790`.

Score and Change, eds. (2020). *Overview of the 2019/2020 La Liga sponsors*. URL: `https://www.scoreandchange.com/overview-2019-2020-la-liga-sponsors/` (visited on 02/15/2021).

Scottish Professional Football League (2021). *The Rules and Regulations of the Scottish Professional Football League*. URL: `https://spfl.co.uk/admin/filemanager/images/shares/pdfs/SPFL%20Rules%20and%20Regulations%2016-Mar-21%20(MASTER%20COPY)%20CLEAN.pdf` (visited on 03/25/2021).

Slowinski, Piper (2012). *Luxury Tax*. URL: `https://library.fangraphs.com/business/luxury-tax/` (visited on 03/21/2021).

Smyth, Rob (2018). *World Cup stunning moments: West Germany 1-0 Austria in 1982*. URL: `https://www.theguardian.com/football/blog/2014/feb/25/world-cup-25-stunning-moments-no3-germany-austria-1982-rob-smyth` (visited on 01/07/2021).

Štrumbelj, E and M Robnik Šikonja (2010). "Online bookmakers' odds as forecasts: The case of European soccer leagues". In: *International Journal of Forecasting* 26.3, pp. 482–488.

Štrumbelj, Erik (2014). "On determining probability forecasts from betting odds". In: *International journal of forecasting* 30.4, pp. 934–943.

Symonds, Tom (2020). *Man denied £1.7m payout by Betfred takes fight to High Court*. URL: `https://www.bbc.co.uk/news/uk-54564536` (visited on 04/13/2021).

The Football Association Premier League Limited, ed. (2019). *Premier League Handbook, Season 2019/20*.

The World Bank (2018). *Gini index (World Bank estimate*. URL: https://data.worldbank.org/indicator/SI.POV.GINI?view=map (visited on 01/12/2021).

UEFA.com (n.d.[a]). *Country Coefficients*. URL: https://www.uefa.com/memberassociations/uefarankings/country/#/yr/2021 (visited on 01/05/2021).

— (n.d.[b]). *Member associations*. URL: https://www.uefa.com/insideuefa/member-associations/ (visited on 01/18/2021).

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

Weisstein, Eric W. (n.d.). *Reflection*. URL: https://mathworld.wolfram.com/Reflection.html (visited on 03/28/2021).

Wesner, Janet (2016). *MAE and RMSE — Which Metric is Better?* URL: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d (visited on 04/21/2021).

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

Wickham, Hadley and Dana Seidel (2020). *scales: Scale Functions for Visualization*. R package version 1.1.1. URL: https://CRAN.R-project.org/package=scales.

Wold, Svante, Kim Esbensen, and Paul Geladi (1987). "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3, pp. 37–52.

Wu, Shaomin (2012). "Warranty data analysis: A review". In: *Quality and Reliability Engineering International* 28.8, pp. 795–805.

# Appendices

# Appendix A

# Definitions

## A.1 Mathematical and Statistical

*Unless stated, these definitions have been taken from the Fifth Edition of the Oxford Concise Dictionary of Mathematics* (Clapham and Nicholson, 2014).

(i) Algorithm — 'A precisely described routine procedure that can be applied and systematically followed through to a conclusion.'

(ii) Central Limit Theorem — '[T]he distribution of the mean of a sequence of random variables tends to a normal distribution as the number in the sequence increases.'

(iii) Cleaning (Data) — Modification, removal, or replacement of 'coarse' ('heaped, censored and missing') data (Wu, 2012).

(iv) Coefficient of Determination, $R^2$ — For a linear model, with data points $x$ associated with a fitted value $f$ and residuals $e_i = y_i - f_i$. Then, $R^2$ is computed by finding the following:

$$\begin{aligned} \text{The mean of the observed data, } \bar{x} &= \tfrac{1}{n}\textstyle\sum_{-1}^{n} x_i \\ \text{Total sum of squares, } SS_{tot} &= \textstyle\sum_i (x_i - \bar{x})^2 \\ \text{Sum of squares of residuals, } SS_{res} &= \textstyle\sum_i e_i^2 \end{aligned}$$

Then,

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

(v) Confidence Interval — 'An interval, calculated from a sample, which contains the value of a certain population parameter with a specified probability.'

(vi) Discrete — '[I]t only takes values from a set of distinct values' (i.e., not continuous).

(vii) Distributions — '[This is] concerned with the way in which the probability of its taking a certain value, or a value within a certain interval, varies. It may be given by the cumulative distribution function[,] its probability mass function [or] its probability density function.'

(viii) Fit — '[T]he degree of of correspondence between the observations and the model's predictions.'

(ix) Kernel Density Estimation — The 'process of finding an estimate probability density function of a random variable. The estimation attempts to infer characteristics of a population, based on a finite data set. The data smoothing problem often is used in signal processing and data science, as it is a powerful way to estimate probability density. In short, the technique allows one to create a smooth curve given a set of random data' (DeepAI, n.d.).

(x) Leverage — The amount of influence each data point $y_i$ can have on each fitted $y$-value, $\hat{y}_j$ (Hoaglin and Welsch, 1978).

(xi) Line of Best Fit — ' A line through a scatter plot of data points that best expresses the relationship between those points' (Chen, 2020).

(xii) Mode — 'For a continuous random variable, [a mode is] a value at which the probability density function has a local maximum.'

(xiii) Normal Distribution — 'The continuous probability distribution wth a probability density function $f$ given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( - \frac{(x-\mu)^2}{2\sigma^2} \right)$$

denoted by $\mathcal{N}o(\mu,\ \sigma^2)$' where $\mu$ is the mean and $\sigma^2$ the variance.

(xiv) Skewness — 'The amount of asymmetry of a distribution... If the distribution has a long tail to the left[...] it is said to be skewed to the left and to have negative skewness.'

(xv) Standard Deviation, $\sigma$ — 'The positive square root of the variance, a commonly used measure of the dispersion of observations in a sample.'

(xvi) Standard Error SE — 'The standard deviation of an estimator of a population parameter.'

(xvii) Variance $\sigma^2$ — '[...] equal to $E[(X-\mu)^2]$,' the expected value of the squared difference between an observation and the mean.

## A.2 Gambling Terms

(i) Bettor — Someone who places a bet. Also *punter*, *gambler*.

(ii) Bookmaker — Organisation that accepts and pays off bets. Also *house*, *bookie*.

(iii) Calciopoli — A match fixing scandal which occurred in the 2004/05 and 2005/06 seasons, where Italian Serie A teams (AC Milan, Fiorentina, Juventus, Lazio and Reggina) 'systematically influenced referees.' (Hafez, 2019).

(iv) Competitive Balance — 'The situation in which no one business of a group of competing businesses has an unfair advantage over the others.' (Cambridge Dictionary, n.d.).

(v) Disgrace of Gijón — A football match between West Germany and Austria at the 1982 FIFA World Cup where a 'mutually suitable scoreline' was played out, ensuring both sides progressed to the knockout rounds, leading to the final pair of subsequent World Cup group stage matches being played simultaneously (Smyth, 2018).

(vi) Favourite-Longshot Bias — An anomaly in betting markets where the favourites win more often then the market probabilities imply; longshots less often (Cain et al., 2000).

(vii) Overround — 'Bookmakers are business people and the prices they offer include a profit margin, which is sometimes referred to as the 'vig' or 'vigorish' in the prices they offer[...]. Overround is [...] defined as the sum of the reciprocals of all prices in a given market. Some definitions of overround subtract 1 from this sum' (MatterOfStats, n.d.). n.b., in this project, overround is defined without this subtraction.

(viii) Stake — The amount of money placed onto a bet by the bettor (the amount of money *at stake*); *the ante*.

# Appendix B

# Chapter 2 Code

```
#### (1) ANALYSING THE ACCURACY OF BETTING ON ELITE EUROPEAN LEAGUES
#Directory, Environment, Packages ----
setwd("~/Desktop/University/University Year 3/331MP/Data")
rm(list=ls())
#If not done before, install these:
#install.packages('car'); install.packages('MASS')
#install.packages('ggplot2'); #install.packages('gridExtra')
library(car); library(MASS); library(ggplot2); library(gridExtra)

### IDA; Analysis of one season, one league [chosen at random] ----
#Reading the data:
fr_l1_1617 <- read.csv("https://www.football-data.co.uk/mmz4281/1617/F1.csv")
fr_l1_1617 <- fr_l1_1617[,c("Div", "Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG", "FTR",
    "BbAvH", "BbAvD", "BbAvA")]
fr_l1_1617 <- na.omit(fr_l1_1617)

#Finding consensus probabilities
#PN => Pre-Normalised (Underlying Probabilities)
fr_l1_1617$AvgHProbPN <- with(fr_l1_1617, round(1/BbAvH, 4))
fr_l1_1617$AvgDProbPN <- with(fr_l1_1617, round(1/BbAvD, 4))
fr_l1_1617$AvgAProbPN <- with(fr_l1_1617, round(1/BbAvA, 4))

fr_l1_1617$Overround <- with(fr_l1_1617, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
fr_l1_1617$AvgHProb <- with(fr_l1_1617, round(AvgHProbPN/Overround,4))
fr_l1_1617$AvgDProb <- with(fr_l1_1617, round(AvgDProbPN/Overround,4))
fr_l1_1617$AvgAProb <- with(fr_l1_1617, round(AvgAProbPN/Overround,4))

#Simple calculations
basic.fr.ida <- matrix(c(mean(fr_l1_1617$AvgHProb), mean(fr_l1_1617$AvgDProb),
    mean(fr_l1_1617$AvgAProb), sd(fr_l1_1617$AvgHProb), sd(fr_l1_1617$AvgDProb),
    sd(fr_l1_1617$AvgAProb)), ncol=3, nrow=2, byrow=T)
rownames(basic.fr.ida) <- c('mean', 'sd')
colnames(basic.fr.ida) <- c('h', 'd', 'a')
basic.fr.ida;round(prop.table(table(fr_l1_1617$FTR)), 4)

```

```
##Histograms
idaHomeHIST <- ggplot(fr_l1_1617, aes(AvgHProb)) + geom_histogram(binwidth=0.05,
    fill="blue") + coord_cartesian(xlim=c(0,1)) + theme_light() + labs(title="Home Win",
    x=NULL, y=NULL)

idaAwayHIST <- ggplot(fr_l1_1617, aes(AvgAProb)) + geom_histogram(binwidth=0.05,
    fill="coral") + coord_cartesian(xlim=c(0,1)) + theme_light() + labs(title="Away Win",
    x=NULL, y=NULL)

idaDrawHIST <- ggplot(fr_l1_1617, aes(AvgDProb)) + geom_histogram(binwidth=0.05,
    fill="green4") + coord_cartesian(xlim=c(0,1)) + theme_light() + labs(title="Draw",
    x=NULL, y=NULL)

ida.histogram <- grid.arrange(idaHomeHIST, idaDrawHIST, idaAwayHIST, nrow=3,
    ncol=1,left="Frequency", bottom="Consensus Probability")

ggsave(path="./writeup/images", filename="elite_01_idahist.png", plot=ida.histogram,
    unit="cm", width=15, height=18)

### EDA ----
##Reading the data using a For Loop:
#Define which countries and seasons we need to read:
countries <- c("de", "en", "es", "fr", "it", "po")
co.we <- c("D1", "E0", "SP1", "F1", "I1", "P1")
#n.b. The Premier League's code is 0; other countries are 1.
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112", "1213", "1314",
    "1415", "1516", "1617", "1718", "1819", "1920")
eliteTemp <- NULL; elite <- NULL
for (i in seasons){
  for (j in 1:6){
    eliteTemp <- read.csv(paste0('https://www.football-data.co.uk/mmz4281/', i, '/',
        co.we[j],'.csv'), fileEncoding = 'latin1')
    eliteTemp$Country <- with(eliteTemp, countries[j])
    eliteTemp$Season <- with(eliteTemp, i)
    if (i=="1920"){
      eliteTemp$BbAvH<-eliteTemp$AvgH; eliteTemp$BbAvA<-eliteTemp$AvgA
      eliteTemp$BbAvD<-eliteTemp$AvgD
    }
    else{}
    eliteTemp <- eliteTemp[ ,c("Div", "Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG",
        "FTR", "BbAvH", "BbAvD", "BbAvA", "Country", "Season")]
    elite <- rbind(elite, eliteTemp)
  }
}
elite <- na.omit(elite)

#Finding underlying probabilities:
#Pre-Normalised Probabilities
elite$AvgHProbPN <- with(elite, round(1/BbAvH, 4))
elite$AvgDProbPN <- with(elite, round(1/BbAvD, 4))
elite$AvgAProbPN <- with(elite, round(1/BbAvA, 4))
#To normalise them:
```

```
74  elite$overround<-with(elite, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
75  elite$AvgHProb <-with(elite, round(AvgHProbPN/overround, 4))
76  elite$AvgDProb <-with(elite, round(AvgDProbPN/overround, 4))
77  elite$AvgAProb <-with(elite, round(AvgAProbPN/overround, 4))
78
79  #For later analysis, we need the Correct/Incorrect Probabilities:
80  N<-nrow(elite); N
81  elite$Correct<-with(elite, rep(0, N))
82  elite$Incorr1<-with(elite, rep(0, N))
83  elite$Incorr2<-with(elite, rep(0, N))
84
85  for (i in 1:N){
86    if ((elite$FTR[i])=="A"){
87      elite$Correct[i]<-(elite$Correct[i] + elite$AvgAProb[i])
88      elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgDProb[i])
89      elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgHProb[i])}
90    else if ((elite$FTR[i])=="H"){
91      elite$Correct[i]<-(elite$Correct[i] + elite$AvgHProb[i])
92      elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgDProb[i])
93      elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgAProb[i])}
94    else {
95      elite$Correct[i]<-(elite$Correct[i] + elite$AvgDProb[i])
96      elite$Incorr1[i]<-(elite$Incorr1[i] + elite$AvgAProb[i])
97      elite$Incorr2[i]<-(elite$Incorr2[i] + elite$AvgHProb[i])}
98  }
99  elite$logCorrect<-with(elite, rep(0,N))
100 for (j in 1:N){elite$logCorrect[j]<-log(elite$Correct[j], base=exp(1))}
101
102 ## Simple calculations
103 basic.elite <- matrix(c(mean(elite$AvgHProb), mean(elite$AvgDProb), mean(elite$AvgAProb),
        sd(elite$AvgHProb), sd(elite$AvgDProb), sd(elite$AvgAProb)), ncol=3, nrow=2, byrow=T)
104 rownames(basic.elite) <- c('mean', 'sd')
105 colnames(basic.elite) <- c('h', 'd', 'a')
106 basic.elite; round(prop.table(table(elite$FTR)), 4) #Observed probabilities
107
108 ### VISUAL ANALYSIS ----
109 ## Boxplots
110 bp.home <- ggplot(elite, aes(x=FTR, y=AvgHProb)) + geom_boxplot(outlier.size=0.75,
        outlier.alpha=0.7, color="blue") + theme_light() + stat_boxplot(coef=5) +
        labs(x="Actual Result", y="Consensus Probability", title = "Home Win", caption = "") +
        coord_cartesian(ylim=c(0,1))
111
112 bp.draw <- ggplot(elite, aes(x=FTR, y=AvgDProb)) + geom_boxplot(outlier.size=0.75,
        outlier.alpha=0.7, color="green4") +theme_light() + stat_boxplot(coef=5) +
        labs(x="Actual Result", y="Consensus Probability", title = "Draw", caption = "") +
        coord_cartesian(ylim=c(0,1))
113
114 bp.away <- ggplot(elite, aes(x=FTR, y=AvgAProb)) + geom_boxplot(outlier.size=0.75,
        outlier.alpha=0.7, color="coral") + theme_light() + stat_boxplot(coef=5) +
        labs(x="Actual Result", y="Consensus Probability", title = "Away Win", caption="Elite
        European Leagues, 2005-2020") + coord_cartesian(ylim=c(0,1))
115
```

```
116 eda.bp.all <- grid.arrange(bp.home, bp.draw, bp.away, nrow=1, ncol=3)
117 ggsave(path="./writeup/images", filename="elite_02_boxplot.png", plot=eda.bp.all,
        unit="cm", width=15, height=10)
118
119 ##Density Plots
120 eda.wins.dens.all <- ggplot(elite, aes(x=AvgHProb, color="HW")) + geom_density() +
        geom_density(data=elite, mapping=aes(x=AvgAProb, color="AW"), show.legend=T) +
        coord_cartesian(xlim=c(0,1)) + labs(title="Home and Away Wins", caption="Elite
        Leagues, 2005-2020", x="Consensus Probability", y="Density") + theme_light() +
        scale_color_manual(name="Market", values=c("HW" = "blue", "AW" = "coral"))
121
122 eda.draw.dens.all <- ggplot(elite, aes(x=AvgDProb, color="D")) + geom_density() +
        coord_cartesian(xlim=c(0,1)) + labs(title="Draws", caption="", x="Consensus
        Probability",y="Density") + theme_light() + scale_color_manual(name = "Market",
        values=c("D" = "green4"))
123
124 eda.density.all <- grid.arrange(eda.wins.dens.all, eda.draw.dens.all, nrow=1, ncol=2,
        left="", bottom="")
125 ggsave(path="./writeup/images", filename="elite_02_edadensall.png", plot=eda.density.all,
        unit="cm", width=15, height=7)
126
127 View(elite[elite$AvgDProb > 0.6,]) #Extremely high P(Draw)
128 View(elite[elite$AvgDProb > 0.35,]) #Unusually high P(Draw)
129
130 #Splitting these by league:
131 #Home Wins
132 eda.home.dens <- ggplot(elite, aes(x=AvgHProb, color=Country)) + geom_density() +
        coord_cartesian(xlim=c(0,1)) + labs(title="Home Win", x=NULL, y=NULL) + theme_light()
        + geom_vline(aes(xintercept=mean(AvgHProb)), linetype="dashed",
        size=0.3)+guides(y="none") + theme(legend.position="none")
133 #Away Wins
134 eda.away.dens <- ggplot(elite, aes(x=AvgAProb, color=Country)) + geom_density() +
        coord_cartesian(xlim=c(0,1)) +labs(title="Away Win", caption = "Elite European
        Leagues, 2005-20", x=NULL, y=NULL) + theme_light() +
        geom_vline(aes(xintercept=mean(AvgAProb)), linetype="dashed", size=0.3)+
        guides(y="none") + theme(legend.position="none")
135 #Draws
136 eda.draw.dens <- ggplot(elite, aes(x=AvgDProb, color=Country)) + geom_density() +
        coord_cartesian(xlim=c(0,0.8)) + labs(title="Draw", x=NULL, y=NULL) +
        geom_vline(aes(xintercept=mean(AvgDProb)), linetype="dashed", size=0.3) +
        guides(y="none") + theme_light() + scale_colour_discrete(labels = c("Germany",
        "England", "Spain", "France", "Italy", "Portugal"))
137
138 eda.density <- grid.arrange(eda.home.dens, eda.draw.dens, eda.away.dens, nrow=3, ncol=1,
        left="", bottom="Consensus Probability")
139 ggsave(path="./writeup/images", filename="elite_02_edadens.png", plot=eda.density,
        unit="cm", width=15, height=18)
140
141 ##Tile Plot
142 #We will group 5+ goals together
143 elite$FTHG.Tile <- with(elite,rep(0,N))
144 elite$FTAG.Tile <- with(elite,rep(0,N))
```

```
for (k in 1:N){
  if ((elite$FTHG[k])>=5){elite$FTHG.Tile[k] <- 5}
  else{elite$FTHG.Tile[k] <- elite$FTHG[k]}}
for (k in 1:N){
  if ((elite$FTAG[k])>=5){elite$FTAG.Tile[k] <- 5}
  else{elite$FTAG.Tile[k] <- elite$FTAG[k]}}

elite.tile <- ggplot(elite, aes(y=FTAG.Tile, x=FTHG.Tile)) + geom_tile(aes(fill =
    Correct)) + scale_fill_distiller(palette = "Greens", direction = 1,
    name="Correct\nProbability") + theme_light() + labs(title="Match Result v. Correct
    Consensus Probability", x="Home Goals Scored", y="Away Goals Scored", caption="Elite
    European Leagues, 2005-2020") + scale_y_discrete(limits=factor(c(1:4, "5+")))
    +scale_x_discrete(limits=factor(c(1:4, "5+"))) +geom_abline(intercept=0, slope=1) +
    coord_cartesian(xlim=c(0,5), ylim=c(0,5))

ggsave(path="./writeup/images", filename="elite_05_tile.png", plot=elite.tile, unit="cm",
    width=15, height=15)
tpbinsizes.elite <- table(elite$FTAG.Tile, elite$FTHG.Tile) #Bin sizes

### CORRELATION ANALYSIS ----

## Binning the data:-
#Home Wins:-
elite$AvgHProb.cut <- cut(elite$AvgHProb, 124, include.lowest=T)
#First, we cut the data into 'bins' choosing 124 breaks
levels(elite$AvgHProb.cut) <- tapply(elite$AvgHProb, elite$AvgHProb.cut, mean)
#Tapply finds the mean of the bin, rather than taking the midpoint
elite.observed.probabilites.TabH <- prop.table(table(elite$FTR, elite$AvgHProb.cut),
    2)[c(1, 2, 3),]
#The c(1,2,3) will remove any extra (blank) rows
elite.observed.probabilites.H <- elite.observed.probabilites.TabH[3,]
#[n,]; if n = : 1 Away; 2 Draw; 3 Home (alphabetic)
elite.bookmaker.probabilites.H <- as.numeric(names(elite.observed.probabilites.H))

#Away Wins:-
elite$AvgAProb.cut <- cut(elite$AvgAProb, 124, inclues.lowest=T)
levels(elite$AvgAProb.cut) <- tapply(elite$AvgAProb, elite$AvgAProb.cut, mean)
elite.observed.probabilites.TabA <- prop.table(table(elite$FTR, elite$AvgAProb.cut),
    2)[c(1, 2, 3), ]
elite.observed.probabilites.A <- elite.observed.probabilites.TabA[1, ]
elite.bookmaker.probabilites.A <- as.numeric(names(elite.observed.probabilites.A))
#Draws:-
elite$AvgDProb.cut <- cut(elite$AvgDProb, 124, inclues.lowest=T)
levels(elite$AvgDProb.cut) <- tapply(elite$AvgDProb, elite$AvgDProb.cut, mean)
elite.observed.probabilites.TabD <- prop.table(table(elite$FTR, elite$AvgDProb.cut),
    2)[c(1, 2, 3), ]
elite.observed.probabilites.D <- elite.observed.probabilites.TabD[2, ]
elite.bookmaker.probabilites.D <- as.numeric(names(elite.observed.probabilites.D))

#Finding R-Squared and RMSE:
elite.bookmaker.probabilities <- c(elite.bookmaker.probabilites.H,
    elite.bookmaker.probabilites.D, elite.bookmaker.probabilites.A)
```

```
186 elite.observed.probabilities <- c(elite.observed.probabilites.H,
        elite.observed.probabilites.D, elite.observed.probabilites.A)
187 #Home Wins
188 elite.lm.home <- lm(elite.observed.probabilites.H ~ elite.bookmaker.probabilites.H)
        #Creates the linear model
189 print(paste("R Squared, Home Win = ", round(summary(elite.lm.home)$r.squared, 5))) #R
        Squared
190 print(paste("RMSE, Home Win = ", round(sqrt(mean(elite.lm.home$residuals^2)), 5))) #RMSE
191 print(paste("Slope, Home Win = ", round(elite.lm.home$coefficients[2], 5))) #Slope
192
193 #Away Wins
194 elite.lm.away <- lm(elite.observed.probabilites.A ~ elite.bookmaker.probabilites.A)
195 print(paste("R Squared, Away Win = ", round(summary(elite.lm.away)$r.squared, 5))) #R
        Squared
196 print(paste("RMSE, Away Win = ", round(sqrt(mean(elite.lm.away$residuals^2)), 5))) #RMSE
197 print(paste("Slope, Away Win = ", round(elite.lm.away$coefficients[2], 5))) #Slope
198
199 #Draws
200 elite.lm.draw <- lm(elite.observed.probabilites.D ~ elite.bookmaker.probabilites.D)
201 print(paste("R Squared, Draw Win = ", round(summary(elite.lm.draw)$r.squared, 5))) #R
        Squared
202 print(paste("RMSE, Draw Win = ", round(sqrt(mean(elite.lm.draw$residuals^2)), 5))) #RMSE
203 print(paste("Slope, Draw Win = ", round(elite.lm.draw$coefficients[2], 5))) #Slope
204
205 #Overall
206 elite.linear.model <- lm(elite.observed.probabilities ~ elite.bookmaker.probabilities)
207 print(paste("R Squared, Overall = ", round(summary(elite.linear.model)$r.squared, 5))) #R
        Squared
208 print(paste("RMSE, Overall = ", round(sqrt(mean(elite.linear.model$residuals^2)), 5)))
        #RMSE
209 print(paste("Slope, Overall = ", round(elite.linear.model$coefficients[2], 5))) #Slope
210
211 #Final Plot
212 elite.scatter <- ggplot(data=NULL,aes()) + geom_smooth() +
213   geom_jitter(aes(x=elite.bookmaker.probabilites.H, y=elite.observed.probabilites.H, color
        = "Home Win"), size=0.75, show.legend=T) +
214   geom_smooth(aes(x=elite.bookmaker.probabilites.H, y=elite.observed.probabilites.H, color
        = "Home Win"), method=lm, alpha=.15, size=0.5) +
215   geom_jitter(aes(x=elite.bookmaker.probabilites.D, y=elite.observed.probabilites.D, color
        = "Draw"), size=0.75, show.legend=T) +
216   geom_smooth(aes(x=elite.bookmaker.probabilites.D, y=elite.observed.probabilites.D, color
        = "Draw"), method=lm, alpha=.15, size=0.5) +
217   geom_jitter(aes(x=elite.bookmaker.probabilites.A, y=elite.observed.probabilites.A, color
        = "Away Win"), size=0.75, show.legend = T) +
218   geom_smooth(aes(x=elite.bookmaker.probabilites.A, y=elite.observed.probabilites.A, color
        = "Away Win"), method=lm, alpha=.15, size=0.5) +
219   geom_abline(intercept = 0, slope = 1, linetype="dashed") + scale_color_manual(name="Bet
        Type", values=c("Home Win" = "blue", "Draw" = "green4", "Away Win" = "coral")) +
220   labs(title="Consensus vs. Observed Probabilities", x="Consensus Probability",
        y="Observed Probability", caption="Elite Euro. Leagues, 2005-2020\nBin Size: 250
        Games") +
221   coord_cartesian(xlim=c(0, 1), ylim=c(0, 1)) + theme_light()
```

```
222
223 ggsave(path = "./writeup/images", filename = "elite_03_scatter.png", plot=elite.scatter,
        unit="cm", width=15, height=10)
224
225 ### HAT VALUES AND LEVERAGE PLOT ----
226 hats <- as.data.frame(hatvalues(elite.lm.draw))
227 leverageCrit <- (2*3)/nrow(hats)
228 hats[hats$`hatvalues(elite.lm.draw)` > leverageCrit,]
229 drawLevPlot <- leveragePlot(elite.lm.draw, elite.bookmaker.probabilites.D, col="green4",
        id = list(method=list(abs(residuals(elite.lm.draw, type="pearson"))), n=10, cex=0.6,
        col="red"), xlab = "Consensus Probability (Draw) | Others", ylab = "Observed
        Probability (Draw) | Others")
230
231 ### PREDICTIVE PERFORMANCE (Overall P1 and P2 Values) ----
232 #Calculating Owen (2009)'s P1 and P2 values for overall (all countries)
233 P1 <- exp( (1/N)*sum(elite$logCorrect) )
234 P2 <- (1/N)*sum( (1-elite$Correct)**2 + (elite$Incorr1)**2 + (elite$Incorr2)**2 )
235
236 ### MODELS FOR EACH LEAGUE AND COUNTRY ----
237 RSqu.H <- NULL; RSqu.D <- NULL; RSqu.A <- NULL; RSqu.O <- NULL
238 RMSE.H <- NULL; RMSE.D <- NULL; RMSE.A <- NULL; RMSE.O <- NULL
239 p1.split <- NULL; p2.split <- NULL
240 slope.H <- NULL; slope.D <- NULL; slope.A <- NULL; slope.O <- NULL
241
242 for(i in countries){
243   modelTempData <- elite[elite$Country==i, ]
244   #Bins
245   modelTempData$AvgHProb.cut <- cut(modelTempData$AvgHProb, 20, include.lowest=T)
246   modelTempData$AvgDProb.cut <- cut(modelTempData$AvgDProb, 5, include.lowest=T)
247   modelTempData$AvgAProb.cut <- cut(modelTempData$AvgAProb, 20, include.lowest=T)
248
249   #Means of each bin
250   levels(modelTempData$AvgHProb.cut) <- tapply(modelTempData$AvgHProb,
          modelTempData$AvgHProb.cut, mean)
251   levels(modelTempData$AvgDProb.cut) <- tapply(modelTempData$AvgDProb,
          modelTempData$AvgDProb.cut, mean)
252   levels(modelTempData$AvgAProb.cut) <- tapply(modelTempData$AvgAProb,
          modelTempData$AvgAProb.cut, mean)
253
254   #Observed Probability for each bin
255   modelTemp.obs.prob.tabH <- prop.table(table(modelTempData$FTR,
          modelTempData$AvgHProb.cut), 2)[c(1, 2, 3), ]
256   modelTemp.obs.prob.tabD <- prop.table(table(modelTempData$FTR,
          modelTempData$AvgDProb.cut), 2)[c(1, 2, 3), ]
257   modelTemp.obs.prob.tabA <- prop.table(table(modelTempData$FTR,
          modelTempData$AvgAProb.cut), 2)[c(1, 2, 3), ]
258
259   modelTemp.obs.prob.H <- modelTemp.obs.prob.tabH[3, ]
260   modelTemp.obs.prob.D <- modelTemp.obs.prob.tabD[2, ]
261   modelTemp.obs.prob.A <- modelTemp.obs.prob.tabA[1, ]
262
263   #Finds the bookmaker probabilities for each group and creates vectors
```

```
264   modelTemp.boo.prob.H <- as.numeric(names(modelTemp.obs.prob.H))
265   modelTemp.boo.prob.D <- as.numeric(names(modelTemp.obs.prob.D))
266   modelTemp.boo.prob.A <- as.numeric(names(modelTemp.obs.prob.A))
267   modelTemp.bookmakers <- c(modelTemp.boo.prob.H, modelTemp.boo.prob.D,
          modelTemp.boo.prob.A)
268   modelTemp.observed <- c(modelTemp.obs.prob.H, modelTemp.obs.prob.D, modelTemp.obs.prob.A)
269
270   #Model creation
271   modelTempH <- lm(modelTemp.obs.prob.H~modelTemp.boo.prob.H)
272   modelTempD <- lm(modelTemp.obs.prob.D~modelTemp.boo.prob.D)
273   modelTempA <- lm(modelTemp.obs.prob.A~modelTemp.boo.prob.A)
274   modelTempO <- lm(modelTemp.observed~modelTemp.bookmakers)
275
276   #Finding values
277   RSqu.H <- c(RSqu.H, round(summary(modelTempH)$r.squared, 5))
278   RMSE.H <- c(RMSE.H, round(sqrt(mean(modelTempH$residuals^2)), 5))
279   slope.H <- c(slope.H, round(modelTempH$coefficients[2], 5))
280   RSqu.D <- c(RSqu.D, round(summary(modelTempD)$r.squared, 5))
281   RMSE.D <- c(RMSE.D, round(sqrt(mean(modelTempD$residuals^2)), 5))
282   slope.D <- c(slope.D, round(modelTempD$coefficients[2], 5))
283   RSqu.A <- c(RSqu.A, round(summary(modelTempA)$r.squared, 5))
284   RMSE.A <- c(RMSE.A, round(sqrt(mean(modelTempA$residuals^2)), 5))
285   slope.A <- c(slope.A, round(modelTempA$coefficients[2], 5))
286   RSqu.O <- c(RSqu.O, round(summary(modelTempO)$r.squared, 5))
287   RMSE.O <- c(RMSE.O, round(sqrt(mean(modelTempO$residuals^2)), 5))
288   slope.O <- c(slope.O, round(modelTempO$coefficients[2], 5))
289
290   p1.temp <- exp((1/(nrow(modelTempData)))*sum(modelTempData$logCorrect))
291   p2.temp <- (1/(nrow(modelTempData))) * sum( (1-modelTempData$Correct)**2 +
          (modelTempData$Incorr1)**2 + (modelTempData$Incorr2)**2 )
292
293   p1.split <- c(p1.split, round(p1.temp, 5))
294   p2.split <- c(p2.split, round(p2.temp, 5))
295 }
296 #Putting this into an easy-to-see Table:
297 league.values <- matrix(c(RSqu.H, RMSE.H, slope.H, RSqu.D, RMSE.D, slope.D, RSqu.A,
        RMSE.A, slope.A, RSqu.O, RMSE.O, slope.O, p1.split, p2.split), ncol=6, byrow=T)
298 rownames(league.values) <- c("RSqu.H", "RMSE.H", "slope.H", "RSqu.D", "RMSE.D", "slope.D",
        "RSqu.A", "RMSE.A", "slope.A", "RSqu.O", "RMSE.O", "slope.O", "p1.split", "p2.split")
299 colnames(league.values) <- countries
300 league.values <- as.table(league.values)
301
302 #For each season:
303 rsqu.season <- NULL; rmse.season <- NULL; p1.season <- NULL; p2.season <- NULL;
        slope.season <- NULL
304 for(i in seasons){
305   modelTempData <- elite[elite$Season==i, ]
306   modelTempData$AvgHProb.cut <- cut(modelTempData$AvgHProb, 10, include.lowest=T)
307   modelTempData$AvgDProb.cut <- cut(modelTempData$AvgDProb, 5, include.lowest=T)
308   modelTempData$AvgAProb.cut <- cut(modelTempData$AvgAProb, 10, include.lowest=T)
309
310   #Finds the mean of each group (cut)
```

```
311   levels(modelTempData$AvgHProb.cut) <- tapply(modelTempData$AvgHProb,
          modelTempData$AvgHProb.cut, mean)
312   levels(modelTempData$AvgDProb.cut) <- tapply(modelTempData$AvgDProb,
          modelTempData$AvgDProb.cut, mean)
313   levels(modelTempData$AvgAProb.cut) <- tapply(modelTempData$AvgAProb,
          modelTempData$AvgAProb.cut, mean)
314
315   #Finds the observed probability for each cut
316   modelTemp.obs.prob.tabH <- prop.table(table(modelTempData$FTR,
          modelTempData$AvgHProb.cut), 2)[c(1, 2, 3), ]
317   modelTemp.obs.prob.tabD <- prop.table(table(modelTempData$FTR,
          modelTempData$AvgDProb.cut), 2)[c(1, 2, 3), ]
318   modelTemp.obs.prob.tabA <- prop.table(table(modelTempData$FTR,
          modelTempData$AvgAProb.cut), 2)[c(1, 2, 3), ]
319
320   modelTemp.obs.prob.H <- modelTemp.obs.prob.tabH[3, ]
321   modelTemp.obs.prob.D <- modelTemp.obs.prob.tabD[2, ]
322   modelTemp.obs.prob.A <- modelTemp.obs.prob.tabA[1, ]
323
324   #Finds the bookmaker probabilities for each group and creates vectors
325   modelTemp.boo.prob.H <- as.numeric(names(modelTemp.obs.prob.H))
326   modelTemp.boo.prob.D <- as.numeric(names(modelTemp.obs.prob.D))
327   modelTemp.boo.prob.A <- as.numeric(names(modelTemp.obs.prob.A))
328   modelTemp.bookmakers <- c(modelTemp.boo.prob.H, modelTemp.boo.prob.D,
          modelTemp.boo.prob.A)
329   modelTemp.observed <- c(modelTemp.obs.prob.H, modelTemp.obs.prob.D, modelTemp.obs.prob.A)
330
331   #Making the model
332   modelTemp0 <- lm(modelTemp.observed~modelTemp.bookmakers)
333   #Finding values
334   rsqu.season <- c(rsqu.season, round(summary(modelTemp0)$r.squared, 5))
335   rmse.season <- c(rmse.season, round(sqrt(mean(modelTemp0$residuals^2)), 5))
336   slope.season <- c(slope.season, round(modelTemp0$coefficients[2], 5))
337
338   p1.temp <- exp((1/(nrow(modelTempData)))*sum(modelTempData$logCorrect))
339   p2.temp <- (1/(nrow(modelTempData))) * sum( (1-modelTempData$Correct)**2 +
          (modelTempData$Incorr1)**2 + (modelTempData$Incorr2)**2 )
340
341   p1.season <- c(p1.season, round(p1.temp, 5))
342   p2.season <- c(p2.season, round(p2.temp, 5))
343 }
344
345 season.values <- matrix(c(rsqu.season, rmse.season, p1.season, p2.season, slope.season),
        ncol=5, byrow=F)
346 colnames(season.values) <- c("rsqu", "rmse", "p1", "p2", "slope")
347 rownames(season.values) <- seasons
348
349 #Plotting Season Values
350 rsqu.se.plot <- ggplot(NULL, aes(y=rsqu.season, x=c(2005:2019))) +
        geom_jitter(color="violetred1")+theme_light() + labs(x = 'Year', y = 'R2') +
        geom_smooth(method = 'lm', color = 'violetred4', se = F)
351 rmse.se.plot <- ggplot(NULL, aes(y=rmse.season, x=c(2005:2019))) +
```

```
    geom_jitter(color="steelblue4")+theme_light() + labs(x = 'Year', y = 'RMSE') +
    geom_smooth(method = 'lm', color = 'steelblue1', se = F)
p1.se.plot <- ggplot(NULL, aes(y=p1.season, x=c(2005:2019))) +
    geom_jitter(color="darkorange4")+theme_light() + labs(x = 'Year', y = 'P1') +
    geom_smooth(method = 'lm', color = 'darkorange1', se = F)
p2.se.plot <- ggplot(NULL, aes(y=p2.season, x=c(2005:2019))) +
    geom_jitter(color="slateblue4")+theme_light() + labs(x = 'Year', y = 'P2') +
    geom_smooth(method = 'lm', color = 'slateblue1', se = F)
slope.se.plot <- ggplot(NULL, aes(y=slope.season, x=c(2005:2019))) +
    geom_jitter(color="springgreen3") + theme_light() + labs(x = "Year", y = "Slope") +
    geom_smooth(method = "lm", color = "springgreen3", se = F) + geom_abline(slope = 0,
    intercept = 1, color = "black") + coord_cartesian(ylim = c(0.5, 1.5))

seasontimeplot <- grid.arrange(rsqu.se.plot, rmse.se.plot, p1.se.plot, p2.se.plot,
    slope.se.plot, ggplot(NULL)+geom_blank()+theme_void(), nrow = 3, top = 'Elite Leagues
    Accuracy Statistics over Time')
ggsave(path="./writeup/images", filename="elite_06_seasontimeplot.png",
    plot=seasontimeplot, unit="cm", width=20, height=20)

### COMPETITIVE BALANCE PCA ----

##LEAGUE MODEL:
#(For leagues, we have the comp. bal. statistics, unlike for seasons)
#We first define the statistics (Gini, NAMSI and K) from Goossens (05):
namsi <- c(0.374, 0.372, 0.364, 0.342, 0.418, 0.505)
kappa <- c(5.71, 5.79, 5.07, 6.00, 5.36, 4.07); invkap <- 1/kappa
gini  <- c(0.723, 0.826, 0.861, 0.784, 0.737, 0.898)

#We define IMBALANCE (Scale (standardise) each statistic above):
namsisc <- scale(namsi); invkapsc <- scale(invkap); ginisc <- scale(gini)
imbalance <- (namsisc[c(1:6),] + invkapsc[c(1:6),] + ginisc[c(1:6),])/3
#The [c(1:6),] cuts off the sd and mean attributes from the scaled data

#We define the LEVEL OF ATTACK - Shots Per Game / Goals Per Game.
attack <- NULL; attackPO <- NULL
for (l in 1:5){
  for (s in seasons){
    dataTemp <- read.csv(paste0("https://www.football-data.co.uk/mmz4281/", s, "/",
        co.we[l], ".csv"))
    dataTemp <- dataTemp[ ,c("FTHG", "FTAG", "HS", "AS")]
    dataTemp$totalGoals <- with(dataTemp, FTHG+FTAG)
    dataTemp$totalShots <- with(dataTemp, HS+AS)
    dataTemp <- na.omit(dataTemp)
    attack <- c(attack, mean(dataTemp$totalShots)/mean(dataTemp$totalGoals))
  }
}
for (s in seasons[13:15]){
  #Data for Po is only avaliable for the 17/18 season onwards.
  dataTemp <- read.csv(paste0("https://www.football-data.co.uk/mmz4281/", s, "/",
      co.we[6], ".csv"))
  dataTemp <- dataTemp[ ,c("FTHG", "FTAG", "HS", "AS")]
  dataTemp$totalGoals <- with(dataTemp, FTHG+FTAG)
```

```
  dataTemp$totalShots <- with(dataTemp, HS+AS)
  attackPO <- c(attackPO, mean(dataTemp$totalShots)/mean(dataTemp$totalGoals))
}
attack <- matrix(attack, ncol=15, byrow = T)
attackPO <- matrix(attackPO, nrow=1, byrow = T)
colnames(attack) <- seasons; rownames(attack) <- countries[1:5]
colnames(attackPO) <- seasons[13:15]; rownames(attackPO) <- countries[6]

attack.league <- c(mean(attack[1,]), mean(attack[2,]), mean(attack[3,]), mean(attack[4,]),
    mean(attack[5,]), mean(attackPO[1,]))
attack.season <- NULL
for (i in 1:12){attack.season <- c(attack.season, mean(attack[,i]))}
for (i in 1:3){attack.season <- c(attack.season, mean(c(attack[1,(i+12)],
    attack[2,(i+12)], attack[3,(i+12)], attack[4,(i+12)], attack[5,(i+12)], attackPO[1,i]
    )))
}

#We define PredAcc, as the normalised sum of the 4 predictive statistics
#We will take the inverse of RMSE and P2
#A high PredAcc value => better bookmaker performance
invrmse.l <- 1/RMSE.O; invp2.l <- 1/p2.split
rsqu.l.sc <- scale(RSqu.O); invrmse.l.sc <- scale(invrmse.l)
p1.l.sc <- scale(p1.split); invp2.l.sc <- scale(invp2.l)
predacc <- (rsqu.l.sc + invrmse.l.sc + p1.l.sc + invp2.l.sc)/4

pc.league <- matrix(c(imbalance, attack.league, predacc), ncol=3, byrow=F)
colnames(pc.league) <- c("imbalance", "attack", "predacc")
rownames(pc.league) <- countries

league.model <- prcomp(pc.league)
league.model$rotation; summary(league.model)

##SEASON MODEL:
pc.season <- matrix(c(rsqu.season, (1/rmse.season), p1.season, (1/p2.season),
    attack.season), ncol = 5, byrow=F)
colnames(pc.season) <- c("rsqu", "inv rmse", "p1", "inv p2", "attack")
rownames(pc.season) <- seasons
pc.season.sc <- scale(pc.season)

season.model <- prcomp(pc.season.sc)
summary(season.model); round(season.model$rotation,3)

##PLOTS:
#League Model Plots
leascree <- ggplot(NULL, aes(x = c(1:3), y = (league.model$sdev)^2)) + geom_line() +
    geom_point(size = 2) + theme_light() + geom_abline(slope = 0, intercept = 1, color =
    "red") +labs(x = "Principal Component", y = "Variances", title = "Screeplot of League
    PCA Components")

leamodlabels <- NULL
for(i in countries){
  if(league.model$x[i,1] > 0){
```

```
435     leamodlabels[i] <- round(league.model$x[i,1],2) + 0.33
436   }
437   else{leamodlabels[i] <- round(league.model$x[i,1],2) - 0.33}
438 }
439
440 leacomps <- ggplot(NULL, aes(countries, league.model$x[,1], color = countries, label =
      round(league.model$x[,1],2))) + geom_pointrange(ymin = 0, ymax = league.model$x[,1]) +
      theme_light() + labs(x = "Country", title = "PC1 Values", y = "Component 1") +
      theme(legend.position = "none") + scale_x_discrete(labels=c("po" = "Portugal", "it" =
      "Italy", "fr" = "France", "es" = "Spain", "en" = "England", "de" = "Germany")) +
      geom_text(aes(y = leamodlabels)) + coord_cartesian(ylim = c(-3,3))
441
442
443 leaguepca <- grid.arrange(leascree, leacomps, ncol = 2, top = "By-League PCA")
444 ggsave(path="./writeup/images", filename="elite_07a_leaguepca.png", plot=leaguepca,
      unit="cm", width=20, height=10)
445
446 #Season Model Plots
447 seascree <- ggplot(NULL, aes(x = c(1:5), y = (season.model$sdev)^2)) + geom_line() +
      geom_point(size = 2) + theme_light() + geom_abline(slope = 0, intercept = 1, color =
      "red") +labs(x = "Principal Component", y = "Variances", title = "Screeplot of Season
      PCA Components")
448
449 seacomps <- ggplot(NULL, aes(x = season.model$x[,1], y = season.model$x[,2], label =
      seasons)) + geom_jitter() + labs(x = 'Component 1', y = 'Component 2', main = "PC1 vs.
      PC2") + theme_light() + geom_text(aes(x = season.model$x[,1], y =
      season.model$x[,2]-0.2))
450
451 seasonpca <- grid.arrange(seascree, seacomps, ncol = 2, top = "By-Season PCA")
452 ggsave(path="./writeup/images", filename="elite_07b_seasonpca.png", plot=seasonpca,
      unit="cm", width=20, height=10)
453
454 #- End -
```

# Appendix C

# Chapter 3 Code

```
#### (2) ANALYSING THE ACCURACY OF BETTING ON ENG/SCO FOOTBALL LEAGUES
#----
#N.B.:
#assume the working directory and libraries are set as in eliteleagues.R
#else run:
#setwd("~/Desktop/University/University Year 3/331MP/Data");rm(list=ls())
library(car); library(MASS); library(ggplot2); library(gridExtra)

### READING DATA ----
#We downloading data straight from football-data.co.uk

divisions <- c("E0", "E1", "E2", "E3", "EC", "SC0", "SC1", "SC2", "SC3")
levels <- c("1", "2", "3")
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112", "1213", "1314",
    "1415", "1516", "1617", "1718", "1819", "1920")

enscoTemp <- NULL; ensco <- NULL
for (i in seasons){
  for (j in divisions){
    enscoTemp <- read.csv(paste0("https://www.football-data.co.uk/mmz4281/", i, "/", j,
        ".csv"), fileEncoding="latin1")
    enscoTemp$Season <- with(enscoTemp, i)
    enscoTemp$Div <- with(enscoTemp, j)
    if (i=="1920"){
      enscoTemp$BbAvH <- enscoTemp$AvgH
      enscoTemp$BbAvA <- enscoTemp$AvgA
      enscoTemp$BbAvD <- enscoTemp$AvgD
      enscoTemp$BbAv.2.5 <- enscoTemp$Avg.2.5
      enscoTemp$BbAv.2.5.1 <- enscoTemp$Avg.2.5.1
      enscoTemp$BbAvAHH <- enscoTemp$AvgAHH
      enscoTemp$BbAvAHA <- enscoTemp$AvgAHA
      enscoTemp$BbAHh <- enscoTemp$AHh}
    else{}
    enscoTemp$Over2.5Odds <- enscoTemp$BbAv.2.5
    enscoTemp$Under2.5Odds <- enscoTemp$BbAv.2.5.1
```

```
    #Greater Than or Less Than don't copy through:
    #A manual check confirms this is the right way round
    enscoTemp$HomeHandicap <- enscoTemp$BbAHh
    enscoTemp <- enscoTemp[,c("Div", "Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG", "FTR",
        "BbAvH", "BbAvD", "BbAvA", "Over2.5Odds", "Under2.5Odds", "HomeHandicap",
        "BbAvAHH", "BbAvAHA", "Season")]
    ensco <- rbind(ensco, enscoTemp)
  }
}
ensco<-na.omit(ensco)

### NORMALISING PROBS, FINDING WINNING BETS ----

#Defining the league 'level': 1 - Elite (EPL, SPL, Championship);
#                             2 - Fully Professional Lower Leagues;
#                             3 - Semi-Professional Lower Leagues.
ensco$Level<-with(ensco, rep(0, nrow(ensco)))
for (k in 1:(nrow(ensco))){
  if (ensco$Div[k]=="E0"){ensco$Level[k]<-1}
  else if (ensco$Div[k]=="E1"){ensco$Level[k]<-1}
  else if (ensco$Div[k]=="E2"){ensco$Level[k]<-2}
  else if (ensco$Div[k]=="E3"){ensco$Level[k]<-2}
  else if (ensco$Div[k]=="EC"){ensco$Level[k]<-3}
  else if (ensco$Div[k]=="SC0"){ensco$Level[k]<-1}
  else if (ensco$Div[k]=="SC1"){ensco$Level[k]<-2}
  else if (ensco$Div[k]=="SC2"){ensco$Level[k]<-3}
  else if (ensco$Div[k]=="SC3"){ensco$Level[k]<-3}
  else{}
}

#Adding and normalising Probability columns
#Pre Normalised:-
  #1X2 Market
    ensco$AvgHProbPN <- with(ensco, round(1/BbAvH, 4))
    ensco$AvgDProbPN <- with(ensco, round(1/BbAvD, 4))
    ensco$AvgAProbPN <- with(ensco, round(1/BbAvA, 4))
  #Under/Over 2.5 goals market
    ensco$Over2.5ProbPN <- with(ensco, round(1/Over2.5Odds, 4))
    ensco$Under2.5ProbPN <- with(ensco, round(1/Under2.5Odds, 4))
  #Asian Handicap Markets
    ensco$AH.HProbPN <- with(ensco, round(1/BbAvAHH, 4))
    ensco$AH.AProbPN <- with(ensco, round(1/BbAvAHA, 4))
#Finding Overrounds:-
    ensco$OneXTwoOverround <- with(ensco, (AvgHProbPN + AvgDProbPN + AvgAProbPN))
    ensco$UnderOverOverround <- with(ensco, (Over2.5ProbPN + Under2.5ProbPN))
    ensco$AHOverround <- with(ensco, (AH.HProbPN + AH.AProbPN))
#Normalising
    ensco$AvgHProb <- with(ensco, round(AvgHProbPN/OneXTwoOverround, 4))
    ensco$AvgDProb <- with(ensco, round(AvgDProbPN/OneXTwoOverround, 4))
    ensco$AvgAProb <- with(ensco, round(AvgAProbPN/OneXTwoOverround, 4))
    ensco$Over2.5Prob <- with(ensco, round(Over2.5ProbPN/UnderOverOverround, 4))
    ensco$Under2.5Prob <- with(ensco, round(Under2.5ProbPN/UnderOverOverround, 4))
```

```
84    ensco$AH.HProb <- with(ensco, round(AH.HProbPN/AHOverround, 4))
85    ensco$AH.AProb <- with(ensco, round(AH.AProbPN/AHOverround, 4))
86
87  #A few important notes:
88  #Rangers had a -275 goal handicap vs. East Fife; assume this meant -2.75:
89  ensco$HomeHandicap[ensco$HomeTeam=="Rangers" & ensco$Date=="11/01/14"] <- -2.75
90  #Hamilton had a 12.5 goal handicap vs. Rangers; assume this meant 1.25:
91  ensco$HomeHandicap[ensco$HomeTeam=="Hamilton" & ensco$Date=="25/10/08"] <- 1.25
92
93  #"Correct" Results -- Straight forward for the 1X2 and U/O Markets:
94  N = nrow(ensco)
95  ensco$Correct1X2 <- with(ensco, rep(0,N))
96  ensco$IncorrectA1X2 <- with(ensco, rep(0,N))
97  ensco$IncorrectB1X2 <- with(ensco, rep(0,N))
98  ensco$TotGoals <- with(ensco, FTHG + FTAG)
99  ensco$CorrectUO <- with(ensco, rep(0,N))
100 ensco$IncorrectUO <- with(ensco, rep(0,N))
101
102 for (l in 1:N){
103   if (ensco$FTR[l] == "H"){
104     ensco$Correct1X2[l] <- ensco$Correct1X2[l] + ensco$AvgHProb[l]
105     ensco$IncorrectA1X2[l] <- ensco$IncorrectA1X2[l] + ensco$AvgAProb[l]
106     ensco$IncorrectB1X2[l] <- ensco$IncorrectB1X2[l] + ensco$AvgDProb[l]}
107   else if (ensco$FTR[l] == "D"){
108     ensco$Correct1X2[l] <- ensco$Correct1X2[l] + ensco$AvgDProb[l]
109     ensco$IncorrectA1X2[l] <- ensco$IncorrectA1X2[l] + ensco$AvgAProb[l]
110     ensco$IncorrectB1X2[l] <- ensco$IncorrectB1X2[l] + ensco$AvgHProb[l]}
111   else if (ensco$FTR[l] == "A"){
112     ensco$Correct1X2[l] <- ensco$Correct1X2[l] + ensco$AvgAProb[l]
113     ensco$IncorrectA1X2[l] <- ensco$IncorrectA1X2[l] + ensco$AvgHProb[l]
114     ensco$IncorrectB1X2[l] <- ensco$IncorrectB1X2[l] + ensco$AvgDProb[l]}
115   else{}
116   if (ensco$TotGoals[l] > 2.5){
117     ensco$CorrectUO[l] <- ensco$CorrectUO[l] + ensco$Over2.5Prob[l]
118     ensco$IncorrectUO[l] <- ensco$IncorrectUO[l] + ensco$Under2.5Prob[l]}
119   else if (ensco$TotGoals[l] < 2.5){
120     ensco$CorrectUO[l] <- ensco$CorrectUO[l] + ensco$Under2.5Prob[l]
121     ensco$IncorrectUO[l] <- ensco$IncorrectUO[l] + ensco$Over2.5Prob[l]}
122   else{}
123 }
124
125 ensco$uo.res <- NULL
126 for (b in 1:N){
127   if (ensco$TotGoals[b] > 2.5){ensco$uo.res[b] <- "over"}
128   else {ensco$uo.res[b] <- "under"}
129 }
130
131 #For Asian Handicaps, we need to work out the winner(s)
132 ensco$FTHG.ah <- with(ensco, rep(0,N))
133 for (m in 1:N){ensco$FTHG.ah[m] <- ensco$FTHG[m] + ensco$HomeHandicap[m]}
134
135 #There are 3 types of AH bets: Assume the bookie bets on the HOME team
```

```
#   Integer: e.g. +1 Handicap for home team
          ##  EVENT                                  #Winner
          # - If there is a draw, or home team wins Home
          # - If away teams by 1 goal              Void (stake refund)
          # - If away team wins by more than 1 goal Away
#   Half: e.g. +0.5 Handicap for the home team
          # - If the home team wins, or it is a draw Home wins
          # - If the away team wins                Away
#   Quarter: e.g. +0.75 Handicap for the home team
#         Half the stake goes to +1, Half goes to +0.5
          # - If the home team wins                Home
          # - If the game is a draw                Home wins
          # - If away team wins by 1 goal          HalfAway wins
          # - If away wins by more than 1 goal     Away wins

ensco$ah.gap <- with(ensco, FTHG.ah - FTAG); ensco$ah.res <- NULL
for (n in 1:N){
  if (ensco$ah.gap[n]<(-0.25)){ensco$ah.res[n]<-"aw"}
  else if (ensco$ah.gap[n]==(-0.25)){ensco$ah.res[n]<-"hfaw"}
  else if (ensco$ah.gap[n]==0){ensco$ah.res[n]<-"vo"}
  else if (ensco$ah.gap[n]==0.25){ensco$ah.res[n]<-"hfhm"}
  else if (ensco$ah.gap[n]>0.25){ensco$ah.res[n]<-"hm"}
  else{}
}

ensco$CorrectAH <- with(ensco, rep(0,N))
ensco$IncorrectAH <- with(ensco, rep(0,N))
#Only considering the FULL wins, rather than half.
for (l in 1:N){
  if (ensco$ah.res[l] == "hm"){
    ensco$CorrectAH[l] <- ensco$CorrectAH[l] + ensco$AH.HProb[l]
    ensco$IncorrectAH[l] <- ensco$IncorrectAH[l] + ensco$AH.AProb[l]}
  else if (ensco$ah.res[l] == "aw"){
    ensco$CorrectAH[l] <- ensco$CorrectAH[l] + ensco$AH.AProb[l]
    ensco$IncorrectAH[l] <- ensco$IncorrectAH[l] + ensco$AH.HProb[l]}
  else{}
}


### BASIC CALCULATIONS ----
#1X2:
basic.1x2 <- NULL
for (a in 1:3){basic.1x2 <- c(basic.1x2, mean(ensco$AvgHProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2, mean(ensco$AvgDProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2, mean(ensco$AvgAProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2, sd(ensco$AvgHProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2, sd(ensco$AvgDProb[ensco$Level == a]))}
for (a in 1:3){basic.1x2 <- c(basic.1x2, sd(ensco$AvgAProb[ensco$Level == a]))}

basic.1x2 <- matrix(c(basic.1x2), ncol=3, byrow=T)
colnames(basic.1x2) <- 1:3
rownames(basic.1x2) <- c("1x2 Home Mean", "1x2 Draw Mean", "1x2 Away Mean", "1x2 Home SD",
```

```
    "1x2 Draw SD", "1x2 Away SD")

#Under/Over:
basic.uo <- NULL
for (a in 1:3){basic.uo <- c(basic.uo, mean(ensco$Under2.5Prob[ensco$Level == a]))}
for (a in 1:3){basic.uo <- c(basic.uo, mean(ensco$Over2.5Prob[ensco$Level == a]))}
for (a in 1:3){basic.uo <- c(basic.uo, sd(ensco$Under2.5Prob[ensco$Level == a]))}
for (a in 1:3){basic.uo <- c(basic.uo, sd(ensco$Over2.5Prob[ensco$Level == a]))}

basic.uo <- matrix(c(basic.uo), ncol=3, byrow=T)
colnames(basic.uo) <- 1:3
rownames(basic.uo) <- c("Under 2.5 Mean", "Over 2.5 Mean", "Under 2.5 SD", "Over 2.5 SD")

#Asian Handicaps:
basic.ah <- NULL
for (a in 1:3){basic.ah <- c(basic.ah, mean(ensco$AH.HProb[ensco$Level == a]))}
for (a in 1:3){basic.ah <- c(basic.ah, mean(ensco$AH.AProb[ensco$Level == a]))}
for (a in 1:3){basic.ah <- c(basic.ah, sd(ensco$AH.HProb[ensco$Level == a]))}
for (a in 1:3){basic.ah <- c(basic.ah, sd(ensco$AH.AProb[ensco$Level == a]))}

basic.ah <- matrix(c(basic.ah), ncol=3, byrow=T)
colnames(basic.ah) <- 1:3
rownames(basic.ah) <- c("AH Home Mean", "AH Away Mean", "AH Home SD", "AH Away SD")

basic.calcs <- round(rbind(basic.1x2, basic.uo, basic.ah),4)

#Observed Probabilities
obsprob.1x2tab <- round(prop.table(table(ensco$FTR, ensco$Level),2), 4)[c(1,2,3), c(1,2,3)]
obsprob.uotab <- round(prop.table(table(ensco$uo.res, ensco$Level),2),4)[c(1,2), c(1,2,3)]
obsprob.ahtab <- round(prop.table(table(ensco$ah.res, ensco$Level),2),4)[c("hm", "hfhm",
    "vo", "hfaw", "aw"),]

#To find the % of AH full wins that are home wins
print(nrow(ensco[ensco$ah.res == "hm",]) / (nrow(ensco[ensco$ah.res == "hm",]) +
    nrow(ensco[ensco$ah.res == "aw",])))

#Finding this by Level
AH.Basic.Proportions <- NULL
for (i in levels){
  tempH <- nrow(ensco[ensco$ah.res == "hm" & ensco$Level == i,])
  tempA <- nrow(ensco[ensco$ah.res == "aw" & ensco$Level == i,])
  tempProp <- tempH / (tempH + tempA)
  AH.Basic.Proportions <- c(AH.Basic.Proportions, tempProp)
}
AH.Basic.Proportions <- as.matrix(AH.Basic.Proportions, nrow = 1, ncol = 3)
rownames(AH.Basic.Proportions) <- paste("Level",levels)
AH.Basic.Proportions

basic.calcs; obsprob.1x2tab; obsprob.uotab; obsprob.ahtab

### PLOTS AND VISUAL ANALYSIS ----
#Density Plots
```

```
237 dens1x2ha <- ggplot(ensco, aes(x=AvgHProb, color="HW")) + geom_density() +
        geom_density(data=ensco, mapping=aes(x=AvgAProb, color="AW")) +
        scale_color_manual(name="Bet Type", values=c("HW" = "blue", "AW" = "coral")) +
        labs(x="Consensus Probabilitiy", y="Density", caption="English and Scottish Leagues,
        2005-2020", title="Home and Away Wins in\nthe 1X2 Market") + theme_light() +
        coord_cartesian(xlim=c(0,1))
238
239 dens1x2d <- ggplot(ensco, aes(x=AvgDProb, color="D")) + geom_density() + labs(x="Consensus
        Probability", y="Density", caption="English and Scottish Leagues, 2005-2020",
        title="\nDraws in \nthe 1X2 Market") + scale_color_manual(name="Bet Type",
        values=c("D" = "green4")) + theme_light() + coord_cartesian(xlim=c(0,1))
240
241 densuo <- ggplot(ensco, aes(x=Under2.5Prob, color="U")) + geom_density() +
        geom_density(data=ensco, aes(x=Over2.5Prob, color="O")) + labs(x="Consensus
        Probability", y="Density", caption="English and Scottish Leagues, 2005-2020",
        title="Under/Over 2.5\nGoals Market") + scale_color_manual(name="Bet Type",
        values=c("U" = "red", "O" = "green")) + theme_light() + coord_cartesian(xlim=c(0,1))
242
243 densah <- ggplot(ensco, aes(x=AH.HProb, color="AH HW")) + geom_density() +
        geom_density(data=ensco, aes(x=AH.AProb, color="AH AW")) + labs(x="Consensus
        Probability", y="Density", caption="English and Scottish Leagues, 2005-2020",
        title="Home and Away Wins in the\nAsian Handicap Market") +
        scale_color_manual(name="Bet Type", values=c("AH HW" = "blue", "AH AW" = "coral")) +
        theme_light() + coord_cartesian(xlim=c(0,1))
244
245 # Saving the plots:
246 ggsave(path = "./writeup/images", filename = "ensco_01_dens1x2ha.png", plot=dens1x2ha,
        unit="cm", width=15, height=10)
247 ggsave(path = "./writeup/images", filename = "ensco_02_dens1x2d.png", plot=dens1x2d,
        unit="cm", width=15, height=10)
248 ggsave(path = "./writeup/images", filename = "ensco_03_densuo.png", plot=densuo,
        unit="cm", width=15, height=10)
249 ggsave(path = "./writeup/images", filename = "ensco_04_densah.png", plot=densah,
        unit="cm", width=15, height=10)
250
251 dens <- grid.arrange(dens1x2ha, dens1x2d, densuo, densah, ncol=2, nrow=2)
252 ggsave(path = "./writeup/images", filename = "ensco_04A_densities.png", plot=dens,
        unit="cm", width=20, height=15)
253
254 #Home Team Handicap v. Mean Consensus P(Home Win)
255 #We'd expect a higher handicap => lower P_cons(Home Win)
256 ensco$AvgHProb.2dp <- with(ensco, round(AvgHProb, 2))
257 ensco$AvgAProb.2dp <- with(ensco, round(AvgAProb, 2))
258
259 ensco$Level <- as.factor(ensco$Level)
260 handicap.v.hprob <- ggplot(ensco, aes(x=AvgHProb.2dp, y=HomeHandicap, color=Level)) +
        geom_count(show.legend = T) + scale_size_area() + theme_light() + labs(x="Consensus
        Probability of a Home Win (1X2)", y="Home Handicap", title="Consensus P(Home Win)\nvs.
        Home Handicap", caption="English and Scottish Leagues, 2005-2020")
261
262 handicap.v.1x2 <- ggplot(ensco, aes(x=AvgAProb.2dp, y=HomeHandicap, color="Away Win")) +
        geom_count(alpha=.5) + scale_size_area() + theme_light() +
```

```
    geom_count(mapping=aes(x=AvgHProb.2dp, y=HomeHandicap, color="Home Win"), alpha=.5) +
    labs(x="Consensus Probability", y="Home Handicap", title="Consensus P(Win, 1X2)\nvs.
    Home Handicap", caption="English and Scottish Leagues, 2005-2020") +
    scale_color_manual(name = "Bet", values = c("Home Win" = "blue", "Away Win" = "coral"))
263
264 ggsave(path = "./writeup/images", filename = "ensco_05_hprob_v_hcap.png",
    plot=handicap.v.hprob, unit="cm", width=15, height=10)
265 ggsave(path = "./writeup/images", filename = "ensco_05a_1x2_v_hcap.png",
    plot=handicap.v.1x2, unit="cm", width=15, height=10)
266
267
268 #Tile Plots for 1X2 and UO Markets
269 #Our highest-scoring draw was 6-6: As before, we bin 6+ (not 5) goals together:
270 N = nrow(ensco); ensco$FTHG.Tile<-with(ensco,rep(0,N))
271 ensco$FTAG.Tile<-with(ensco,rep(0,N))
272
273 for (k in 1:N){
274   if ((ensco$FTHG[k])>=6){ensco$FTHG.Tile[k]<-6}
275   else{ensco$FTHG.Tile[k]<-ensco$FTHG[k]}}
276 for (k in 1:N){
277   if ((ensco$FTAG[k])>=6){ensco$FTAG.Tile[k]<-6}
278   else{ensco$FTAG.Tile[k]<-ensco$FTAG[k]}}
279
280 tile.1x2 <- ggplot(ensco, aes(y=FTAG.Tile, x=FTHG.Tile)) + geom_tile(aes(fill =
    Correct1X2)) + scale_fill_distiller(palette = "Greens", direction = 1, name="Correct
    1X2\nProbability") + theme_light() + labs(title="Result vs. Correct
    Consensus\nProbability in the 1X2 Market", x="Home Goals", y="Away Goals",
    caption="English and Scottish Leagues, 2005-2020") +
    scale_y_discrete(limits=factor(c(1:5, "6+"))) + scale_x_discrete(limits=factor(c(1:5,
    "6+"))) + geom_abline(intercept=0, slope=1) + coord_cartesian(xlim=c(0,6), ylim=c(0,6))
281
282 tile.uo <- ggplot(ensco, aes(y=FTAG.Tile, x=FTHG.Tile)) + geom_tile(aes(fill =
    Over2.5Prob)) + scale_fill_distiller(palette = "Paired", direction = 1, name="P(Over
    2.5 Goals)") + theme_light() + labs(title="Result vs. Consensus\nProbability of Over
    2.5 Goals", x="Home Goals", y="Away Goals", caption="English and Scottish Leagues,
    2005-2020") + scale_y_discrete(limits=factor(c(1:5, "6+"))) +
    scale_x_discrete(limits=factor(c(1:5, "6+"))) + coord_cartesian(xlim=c(0,6),
    ylim=c(0,6)) + geom_segment(aes(x=2.5,xend=2.5,y=-5,yend=2.5),color="black") +
    geom_segment(aes(x=-5,xend=2.5,y=2.5,yend=2.5),color="black")
283
284 ggsave(path = "./writeup/images", filename = "ensco_06_tile_1x2.png", plot=tile.1x2,
    unit="cm", width=15, height=10)
285 ggsave(path = "./writeup/images", filename = "ensco_07_tile_uo.png", plot=tile.uo,
    unit="cm", width=15, height=10)
286 tpbinsizes.ensco <- table(ensco$FTAG.Tile, ensco$FTHG.Tile) #Bin sizes
287
288 ##Under/Over v. Handicap
289 #Are higher handicap games (more 'obvious') likely to have higher goals?
290 ensco$Over2.5Prob.2dp <- with(ensco, round(Over2.5Prob, 2))
291 handicap.v.over2.5 <- ggplot(ensco, aes(x=Over2.5Prob.2dp, y=HomeHandicap)) + geom_count()
    + scale_size_area() + theme_light() + labs(x="Consensus Probability of a Over 2.5
    Goals", y="Home Handicap", title="Consensus P(Over 2.5 Goals)\nvs. Handicap",
```

```
    caption="English and Scottish Leagues, 2005-2020")

ensco$GoalDiffTile <- with(ensco, FTAG.Tile-FTHG.Tile)
expected.v.act.difference <- ggplot(ensco, aes(y=GoalDiffTile, x=HomeHandicap)) +
    geom_count() + scale_size_area() + theme_light() + labs(x="Home Handicap (Expected
    GD)", y="Away Goals minus Home Goals (Actual GD)", title="Expected vs. Actual Goal
    Difference", caption="English and Scottish Leagues, 2005-2020")

ggsave(path = "./writeup/images", filename = "ensco_08_handicap_v_over.png",
    plot=handicap.v.over2.5, unit="cm", width=15, height=10)
ggsave(path = "./writeup/images", filename = "ensco_09_exp_v_act_goaldiff.png",
    plot=expected.v.act.difference, unit="cm", width=15, height=10)

### CORRELATION TESTS (Kendall's Tau, Spearman) ----
cor.test(ensco$FTHG.Tile, ensco$FTAG.Tile, method = "kendall")
cor.test(ensco$FTHG.Tile, ensco$FTAG.Tile, method = 'spearman')
#p << 0.005: Strong evidence of an association
library(DescTools)
GoodmanKruskalGamma(ensco$FTHG.Tile, ensco$FTAG.Tile, conf.level = 0.95)

### CORRELATION ANALYSIS: MODEL CREATION (Overall) ----

#Cutting and defining the levels:
ensco$AvgHProb.cut <- cut(ensco$AvgHProb, 50, include.lowest=T)
levels(ensco$AvgHProb.cut) <- tapply(ensco$AvgHProb, ensco$AvgHProb.cut, mean)
ensco$AvgDProb.cut <- cut(ensco$AvgDProb, 50, include.lowest=T)
levels(ensco$AvgDProb.cut) <- tapply(ensco$AvgDProb, ensco$AvgDProb.cut, mean)
ensco$AvgAProb.cut <- cut(ensco$AvgAProb, 50, include.lowest=T)
levels(ensco$AvgAProb.cut) <- tapply(ensco$AvgAProb, ensco$AvgAProb.cut, mean)

ensco$Over2.5Prob.cut <- cut(ensco$Over2.5Prob, 50, include.lowest=T)
levels(ensco$Over2.5Prob.cut) <- tapply(ensco$Over2.5Prob, ensco$Over2.5Prob.cut, mean)

ensco$AH.HProb.cut <- cut(ensco$AH.HProb, 50, include.lowest=T)
levels(ensco$AH.HProb.cut) <- tapply(ensco$AH.HProb, ensco$AH.HProb.cut, mean)
ensco$AH.AProb.cut <- cut(ensco$AH.AProb, 50, include.lowest=T)
levels(ensco$AH.AProb.cut) <- tapply(ensco$AH.AProb, ensco$AH.AProb.cut, mean)

#Observed Probability for each cut:
obsprob.1x2.H <- prop.table(table(ensco$FTR, ensco$AvgHProb.cut), 2)[3,]
booprob.1x2.H <- as.numeric(names(obsprob.1x2.H))
obsprob.1x2.D <- prop.table(table(ensco$FTR, ensco$AvgDProb.cut), 2)[2,]
booprob.1x2.D <- as.numeric(names(obsprob.1x2.D))
obsprob.1x2.A <- prop.table(table(ensco$FTR, ensco$AvgAProb.cut), 2)[1,]
booprob.1x2.A <- as.numeric(names(obsprob.1x2.A))

booprob.1x2 <- c(booprob.1x2.H,booprob.1x2.D,booprob.1x2.A)
obsprob.1x2 <- c(obsprob.1x2.H,obsprob.1x2.D,obsprob.1x2.A)

obsprob.uo <- prop.table(table(ensco$uo.res, ensco$Over2.5Prob.cut), 2)[1,]
booprob.uo <- as.numeric(names(obsprob.uo))

```

```
#For AH bets, we will only take full wins:
obsprob.ah.H <- prop.table(table(ensco$ah.res, ensco$AH.HProb.cut), 2)[4,] #4 = Home
booprob.ah.H <- as.numeric(names(obsprob.ah.H))
obsprob.ah.A <- prop.table(table(ensco$ah.res, ensco$AH.AProb.cut), 2)[1,] #1 = Away
booprob.ah.A <- as.numeric(names(obsprob.ah.A))

booprob.ah <- c(booprob.ah.H, booprob.ah.A)
obsprob.ah <- c(obsprob.ah.H, obsprob.ah.A)

#Final models
model.1x2.h <- lm(obsprob.1x2.H~booprob.1x2.H)
model.1x2.d <- lm(obsprob.1x2.D~booprob.1x2.D)
model.1x2.a <- lm(obsprob.1x2.A~booprob.1x2.A)
model.1x2.o <- lm(obsprob.1x2~booprob.1x2)
model.uo <- lm(obsprob.uo~booprob.uo)
model.ah <- lm(obsprob.ah~booprob.ah)

#R Squared and RMSE values:-
rsrm.val.1x2 <- matrix(c(summary(model.1x2.h)$r.squared, summary(model.1x2.d)$r.squared,
    summary(model.1x2.a)$r.squared, sqrt(mean(model.1x2.h$residuals^2)),
    sqrt(mean(model.1x2.d$residuals^2)), sqrt(mean(model.1x2.a$residuals^2)) ), ncol = 3,
    nrow = 2, byrow=T, dimnames = list(c("RSq", "RMSE"), c("1X2: H", "D", "A")))
rsrm.val.uoah <- matrix(c(summary(model.uo)$r.squared, summary(model.ah)$r.squared,
    sqrt(mean(model.uo$residuals^2)), sqrt(mean(model.ah$residuals^2)) ), ncol = 2, nrow =
    2, byrow=T, dimnames = list(c("RSq", "RMSE"), c("Under/Over", "AH")))

#These show that the odds for the U/O market, 1x2 Draws are NOT accurate

#    Model Plots ----
convobs.1x2 <- ggplot(data=NULL, aes()) + geom_smooth() +
  geom_jitter(aes(x=booprob.1x2.H, y=obsprob.1x2.H, color="1X2 Home"), size=0.75) +
  geom_smooth(aes(x=booprob.1x2.H, y=obsprob.1x2.H, color="1X2 Home"), method=lm) +
  geom_jitter(aes(x=booprob.1x2.D, y=obsprob.1x2.D, color="1X2 Draw"), size=0.75) +
  geom_smooth(aes(x=booprob.1x2.D, y=obsprob.1x2.D, color="1X2 Draw"), method=lm) +
  geom_jitter(aes(x=booprob.1x2.A, y=obsprob.1x2.A, color="1X2 Away"), size=0.75) +
  geom_smooth(aes(x=booprob.1x2.A, y=obsprob.1x2.A, color="1X2 Away"), method=lm) +
  geom_abline(intercept=0, slope=1, linetype="dashed") + theme_light() +
  labs(x = "Bookmaker Consensus Probability", y = "Observed Probability", caption="Eng/Sco
      05-20") + scale_color_manual(name="Bet Type", values=c("1X2 Home" = "blue", "1X2
      Draw" = "green4", "1X2 Away" = "coral")) + coord_cartesian(xlim=c(0,1), ylim=c(0,1))

convobs.uo <- ggplot(data=NULL, aes()) + geom_smooth() +
  geom_jitter(aes(x=booprob.uo, y=obsprob.uo, color="Under/Over"), size=0.75) +
  geom_smooth(aes(x=booprob.uo, y=obsprob.uo, color="Under/Over"), method=lm) +
  geom_abline(intercept=0, slope=1, linetype="dashed") + theme_light() +
  labs(x = "Bookmaker Consensus Probability", y = "Observed Probability", caption="Eng/Sco
      05-20") + scale_color_manual(name="Bet Type", values=c("Under/Over" = "red")) +
      coord_cartesian(xlim=c(0,1), ylim=c(0,1))

convobs.ah <- ggplot(data=NULL, aes()) + geom_smooth() +
  geom_jitter(aes(x=booprob.ah.H, y=obsprob.ah.H, color="AH Home"), size=0.75) +
  geom_smooth(aes(x=booprob.ah.H, y=obsprob.ah.H, color="AH Home"), method=lm) +
```

```
  geom_jitter(aes(x=booprob.ah.A, y=obsprob.ah.A, color="AH Away"), size=0.75) +
  geom_smooth(aes(x=booprob.ah.A, y=obsprob.ah.A, color="AH Away"), method=lm) +
  geom_abline(intercept=0, slope=1, linetype="dashed") + theme_light() +
  labs(x = "Bookmaker Consensus Probability", y = "Observed Probability", caption="Eng/Sco
      05-20") + scale_color_manual(name="Bet Type", values=c("AH Home" = "blue", "AH Away"
      = "coral")) + coord_cartesian(xlim=c(0,1), ylim=c(0,1))

ggsave(path = "./writeup/images", filename = "ensco_10_convobs_1x2.png", plot=convobs.1x2,
    unit="cm", width=15, height=6)
ggsave(path = "./writeup/images", filename = "ensco_11_convobs_uo.png", plot=convobs.uo,
    unit="cm", width=15, height=6)
ggsave(path = "./writeup/images", filename = "ensco_12_convobs_ah.png", plot=convobs.ah,
    unit="cm", width=15, height=6)

### CORRELATION ANALYSIS: PER LEVEL ----
#To view sample size:-
for (j in levels){
  DataTemp <- ensco[ensco$Level==j,]
  print(paste0("For level ",j,", n = ",nrow(DataTemp)))
}
rsqu.level <- NULL; rmse.level <- NULL; rsqu.level.1x2 <- NULL; rmse.level.1x2 <- NULL;
rsqu.level.uo <- NULL; rmse.level.uo <- NULL; rsqu.level.ah <- NULL; rmse.level.ah <- NULL;
p1.level <- NULL; p2.level <- NULL; p1.uo.level <- NULL; p2.uo.level <- NULL; p1.ah.level
    <- NULL; p2.ah.level <- NULL
slope.level.1x2 <- NULL; slope.level.uo <- NULL; slope.level.ah <- NULL

ensco$LogCorrect1x2 <- with(ensco, log(ensco$Correct1X2))
ensco$LogCorrectUO <- with(ensco, log(ensco$CorrectUO))

#As we're taking the log and results with AH voids/half-wins, we create a subset:
ensco.ah.results <- ensco[ensco$CorrectAH > 0,]
ensco.ah.results$LogCorrectAH <- with(ensco.ah.results, log(ensco.ah.results$CorrectAH))

#n.b. We don't do a model for each -- just an overall 1x2, AH and U/O.
#     P1 and P2 is based purely on the 1X2 market.

for (j in levels){
  dataTemp <- ensco[ensco$Level==j,]
  dataTempAH <- ensco.ah.results[ensco.ah.results$Level==j,] #For AH P Values
  dataTemp$AvgHProb.cut <- cut(dataTemp$AvgHProb, 35, include.lowest = T)
  levels(dataTemp$AvgHProb.cut) <- tapply(dataTemp$AvgHProb, dataTemp$AvgHProb.cut, mean)
  dataTemp$AvgDProb.cut <- cut(dataTemp$AvgDProb, 15, include.lowest = T)
  levels(dataTemp$AvgDProb.cut) <- tapply(dataTemp$AvgDProb, dataTemp$AvgDProb.cut, mean)
  dataTemp$AvgAProb.cut <- cut(dataTemp$AvgAProb, 35, include.lowest = T)
  levels(dataTemp$AvgAProb.cut) <- tapply(dataTemp$AvgAProb, dataTemp$AvgAProb.cut, mean)

  dataTemp$Over2.5Prob.cut <- cut(dataTemp$Over2.5Prob, 35, include.lowest = T)
  levels(dataTemp$Over2.5Prob.cut) <- tapply(dataTemp$Over2.5Prob,
      dataTemp$Over2.5Prob.cut, mean)

  dataTemp$AH.HProb.cut <- cut(dataTemp$AH.HProb, 35, include.lowest = T)
  levels(dataTemp$AH.HProb.cut) <- tapply(dataTemp$AH.HProb, dataTemp$AH.HProb.cut, mean)
```

```
426 dataTemp$AH.AProb.cut <- cut(dataTemp$AH.AProb, 35, include.lowest = T)
427 levels(dataTemp$AH.AProb.cut) <- tapply(dataTemp$AH.AProb, dataTemp$AH.AProb.cut, mean)
428
429 obs.1x2.h <- prop.table(table(dataTemp$FTR, dataTemp$AvgHProb.cut), 2)[3,]
430 obs.1x2.d <- prop.table(table(dataTemp$FTR, dataTemp$AvgDProb.cut), 2)[2,]
431 obs.1x2.a <- prop.table(table(dataTemp$FTR, dataTemp$AvgAProb.cut), 2)[1,]
432 obs.1x2 <- c(obs.1x2.h, obs.1x2.d, obs.1x2.a)
433
434 boo.1x2.h <- as.numeric(names(obs.1x2.h))
435 boo.1x2.d <- as.numeric(names(obs.1x2.d))
436 boo.1x2.a <- as.numeric(names(obs.1x2.a))
437 boo.1x2 <- c(boo.1x2.h, boo.1x2.d, boo.1x2.a)
438
439 obs.uo <- prop.table(table(dataTemp$uo.res, dataTemp$Over2.5Prob.cut), 2)[1,]
440 boo.uo <- as.numeric(names(obs.uo))
441
442 obs.ah.h <- prop.table(table(dataTemp$ah.res, dataTemp$AH.HProb.cut), 2)[4,]
443 obs.ah.a <- prop.table(table(dataTemp$ah.res, dataTemp$AH.AProb.cut), 2)[1,]
444 obs.ah <- c(obs.ah.h, obs.ah.a)
445
446 boo.ah.h <- as.numeric(names(obs.ah.h))
447 boo.ah.a <- as.numeric(names(obs.ah.a))
448 boo.ah <- c(boo.ah.h, boo.ah.a)
449
450 modelTemp.1x2 <- lm(obs.1x2 ~ boo.1x2)
451 modelTemp.uo <- lm(obs.uo ~ boo.uo)
452 modelTemp.ah <- lm(obs.ah ~ boo.ah)
453 par(mfrow=c(3,1))
454 plot(modelTemp.1x2, 5, main = paste0("1X2 Market; Level ",j))
455 plot(modelTemp.uo, 5, main = paste0("UO Market; Level ",j))
456 plot(modelTemp.ah, 5, main = paste0("AH Market; Level ",j))
457 par(mfrow=c(1,1))
458
459 plot.1x2 <- ggplot(NULL, aes(x=boo.1x2.h, y=obs.1x2.h, color="Home")) +
460   geom_smooth(method="lm", alpha=0.3) +
461   geom_smooth(aes(x=boo.1x2.a, y=obs.1x2.a, color="Away"), method="lm", alpha=0.3) +
462   geom_smooth(aes(x=boo.1x2.d, y=obs.1x2.d, color="Draw"), method="lm", alpha=0.3) +
463   geom_jitter(aes(color="Home"), shape=1) +
464   geom_jitter(aes(x=boo.1x2.a, y=obs.1x2.a, color="Away"), shape=1) +
465   geom_jitter(aes(x=boo.1x2.d, y=obs.1x2.d, color="Draw"), shape=1) +
466   geom_abline(slope=1, intercept=0, color="black", linetype="dashed") +
467   coord_cartesian(xlim=c(0,1), ylim=c(0,1)) + labs(x = paste0("1X2 Bookmaker Consensus
        Probabilities: Level ",j), y = NULL) + scale_color_manual(name="Bet Type", values
        = c("Home" = "blue", "Away" = "coral", "Draw" = "green4")) + theme_light()
468
469 plot.uo <- ggplot(NULL, aes(x=boo.uo, y=obs.uo, color="Over 2.5 Goals")) +
470   geom_smooth(method="lm", alpha=0.3) +
471   geom_jitter(shape=2) +
472   geom_abline(slope=1, intercept=0, color="black", linetype="dashed") +
473   coord_cartesian(xlim=c(0,1), ylim=c(0,1)) + labs(x=paste0("UO Bookmaker Consensus
        Probabilities: Level ",j), y=NULL) + scale_color_manual(name="Bet Type",
        values=c("Over 2.5 Goals" = "red")) + theme_light()
```

```
474
475   plot.ah <- ggplot(NULL, aes(x=boo.ah.h, y=obs.ah.h, color="AH Home")) +
476     geom_smooth(method = "lm", alpha=0.3) +
477     geom_smooth(aes(x=boo.ah.a, y=obs.ah.a, color="AH Away"), method="lm", alpha=0.3) +
          geom_jitter(shape=5) +
478     geom_jitter(aes(x=boo.ah.a, y=obs.ah.a, color="AH Away"), method="lm", alpha=0.3) +
479     geom_abline(slope=1, intercept=0, color="black", linetype="dashed") +
480     coord_cartesian(xlim=c(0,1), ylim=c(0,1)) + labs(x= paste0("AH Bookmaker Consensus
          Probabilities: Level ",j), y=NULL) + scale_color_manual(name="Bet Type",
          values=c("AH Home"="blue", "AH Away"="coral")) + theme_light()
481
482
483   plotTemp <- grid.arrange(plot.1x2, plot.uo, plot.ah, nrow=3, ncol=1, left="Observed
          Probability")
484   ggsave(path = "./writeup/images", filename = paste0("ensco_13_level",j,".png"),
          plot=plotTemp, unit="cm", width=15, height=10)
485
486
487   rsqu.level.1x2 <- c(rsqu.level.1x2, round(summary(modelTemp.1x2)$r.squared, 5))
488   rmse.level.1x2 <- c(rmse.level.1x2, round(sqrt(mean(modelTemp.1x2$residuals^2)), 5))
489   rsqu.level.uo <- c(rsqu.level.uo, round(summary(modelTemp.uo)$r.squared, 5))
490   rmse.level.uo <- c(rmse.level.uo, round(sqrt(mean(modelTemp.uo$residuals^2)), 5))
491   rsqu.level.ah <- c(rsqu.level.ah, round(summary(modelTemp.ah)$r.squared, 5))
492   rmse.level.ah <- c(rmse.level.ah, round(sqrt(mean(modelTemp.ah$residuals^2)), 5))
493
494   p1.temp <- exp(1/(nrow(dataTemp)) * sum(dataTemp$LogCorrect1x2))
495   p2.temp <- 1/(nrow(dataTemp))*sum((1 - dataTemp$Correct1X2)**2 +
          (dataTemp$IncorrectA1X2)**2 + (dataTemp$IncorrectB1X2)**2 )
496   p1.level <- c(p1.level, round(p1.temp, 5))
497   p2.level <- c(p2.level, round(p2.temp ,5))
498
499   p1.uo.temp <- exp(1/(nrow(dataTemp)) * sum(dataTemp$LogCorrectUO))
500   p2.uo.temp <- 1/(nrow(dataTemp))*sum((1 - dataTemp$CorrectUO)**2 +
          (dataTemp$IncorrectUO)**2)
501   p1.uo.level <- c(p1.uo.level, round(p1.uo.temp, 5))
502   p2.uo.level <- c(p2.uo.level, round(p2.uo.temp ,5))
503
504   p1.ah.temp <- exp(1/(nrow(dataTempAH)) * sum(dataTempAH$LogCorrectAH))
505   p2.ah.temp <- 1/(nrow(dataTempAH)) * sum((1 - dataTempAH$CorrectAH)**2 +
          (dataTempAH$IncorrectAH)**2)
506   p1.ah.level <- c(p1.ah.level, round(p1.ah.temp, 5))
507   p2.ah.level <- c(p2.ah.level, round(p2.ah.temp, 5))
508
509   slope.level.1x2 <- c(slope.level.1x2, modelTemp.1x2$coefficients[2]) #Coefficients[2] is
          the gradient
510   slope.level.uo <- c(slope.level.uo, modelTemp.uo$coefficients[2])
511   slope.level.ah <- c(slope.level.ah, modelTemp.ah$coefficients[2])
512
513 }
514 rmse.level <- matrix( c(rmse.level.1x2, rmse.level.uo, rmse.level.ah), ncol=3, byrow=T,
      dimnames = list(c("1x2", "UO", "AH"), levels))
515 rsqu.level <- matrix( c(rsqu.level.1x2, rsqu.level.uo, rsqu.level.ah), ncol=3, byrow=T,
```

```
    dimnames = list(c("1x2", "UO", "AH"), levels))
p.values.level <- matrix(c(p1.level, p2.level, p1.uo.level, p2.uo.level, p1.ah.level,
    p2.ah.level), ncol=3, byrow=T, dimnames = list(c("P1 1X2", "P2 1X2", "P1 UO", "P2 UO",
    "P1 AH", "P2 AH"), levels))
slope.level <- matrix(c(slope.level.1x2, slope.level.uo, slope.level.ah), ncol = 3, byrow
    = F, dimnames = list(levels, c("1X2", "UO", "AH")))

# ---- Comparing Seasons

rsqu.season.1x2 <- NULL; rmse.season.1x2 <- NULL; p1.season.1x2 <- NULL; p2.season.1x2 <-
    NULL
rsqu.season.uo <- NULL; rmse.season.uo <- NULL; p1.season.uo <- NULL; p2.season.uo <- NULL
rsqu.season.ah <- NULL; rmse.season.ah <- NULL; p1.season.ah <- NULL; p2.season.ah <- NULL
slope.season.1x2 <- NULL; slope.season.uo <- NULL; slope.season.ah <- NULL

for(i in seasons){
  dataTemp <- ensco[ensco$Season == i, ]

  dataTemp$AvgHProb.cut <- cut(dataTemp$AvgHProb, 10, include.lowest = T)
  levels(dataTemp$AvgHProb.cut) <- tapply(dataTemp$AvgHProb, dataTemp$AvgHProb.cut, mean)
  dataTemp$AvgDProb.cut <- cut(dataTemp$AvgDProb, 5, include.lowest = T)
  levels(dataTemp$AvgDProb.cut) <- tapply(dataTemp$AvgDProb, dataTemp$AvgDProb.cut, mean)
  dataTemp$AvgAProb.cut <- cut(dataTemp$AvgAProb, 10, include.lowest = T)
  levels(dataTemp$AvgAProb.cut) <- tapply(dataTemp$AvgAProb, dataTemp$AvgAProb.cut, mean)

  dataTemp$Over2.5Prob.cut <- cut(dataTemp$Over2.5Prob, 10, include.lowest = T)
  levels(dataTemp$Over2.5Prob.cut) <- tapply(dataTemp$Over2.5Prob,
      dataTemp$Over2.5Prob.cut, mean)

  dataTemp$AH.HProb.cut <- cut(dataTemp$AH.HProb, 10, include.lowest = T)
  levels(dataTemp$AH.HProb.cut) <- tapply(dataTemp$AH.HProb, dataTemp$AH.HProb.cut, mean)
  dataTemp$AH.AProb.cut <- cut(dataTemp$AH.AProb, 10, include.lowest = T)
  levels(dataTemp$AH.AProb.cut) <- tapply(dataTemp$AH.AProb, dataTemp$AH.AProb.cut, mean)

  obs.1x2.h <- prop.table(table(dataTemp$FTR, dataTemp$AvgHProb.cut), 2)[3,]
  obs.1x2.d <- prop.table(table(dataTemp$FTR, dataTemp$AvgDProb.cut), 2)[2,]
  obs.1x2.a <- prop.table(table(dataTemp$FTR, dataTemp$AvgAProb.cut), 2)[1,]
  obs.1x2 <- c(obs.1x2.h, obs.1x2.d, obs.1x2.a)

  boo.1x2.h <- as.numeric(names(obs.1x2.h))
  boo.1x2.d <- as.numeric(names(obs.1x2.d))
  boo.1x2.a <- as.numeric(names(obs.1x2.a))
  boo.1x2 <- c(boo.1x2.h, boo.1x2.d, boo.1x2.a)

  obs.uo <- prop.table(table(dataTemp$uo.res, dataTemp$Over2.5Prob.cut), 2)[1,]
  boo.uo <- as.numeric(names(obs.uo))

  obs.ah.h <- prop.table(table(dataTemp$ah.res, dataTemp$AH.HProb.cut), 2)[4,]
  obs.ah.a <- prop.table(table(dataTemp$ah.res, dataTemp$AH.AProb.cut), 2)[1,]
  obs.ah <- c(obs.ah.h, obs.ah.a)

  boo.ah.h <- as.numeric(names(obs.ah.h))
```

```
562   boo.ah.a <- as.numeric(names(obs.ah.a))
563   boo.ah <- c(boo.ah.h, boo.ah.a)
564
565   modelTemp.1x2 <- lm(obs.1x2 ~ boo.1x2)
566   modelTemp.uo <- lm(obs.uo ~ boo.uo)
567   modelTemp.ah <- lm(obs.ah ~ boo.ah)
568
569   rsqu.season.1x2 <- c(rsqu.season.1x2, round(summary(modelTemp.1x2)$r.squared, 5))
570   rmse.season.1x2 <- c(rmse.season.1x2, round(sqrt(mean(modelTemp.1x2$residuals^2)), 5))
571   rsqu.season.uo <- c(rsqu.season.uo, round(summary(modelTemp.uo)$r.squared, 5))
572   rmse.season.uo <- c(rmse.season.uo, round(sqrt(mean(modelTemp.uo$residuals^2)), 5))
573   rsqu.season.ah <- c(rsqu.season.ah, round(summary(modelTemp.ah)$r.squared, 5))
574   rmse.season.ah <- c(rmse.season.ah, round(sqrt(mean(modelTemp.ah$residuals^2)), 5))
575
576   p1.temp <- exp(1/(nrow(dataTemp)) * sum(dataTemp$LogCorrect1x2))
577   p2.temp <- 1/(nrow(dataTemp))*sum((1 - dataTemp$Correct1X2)**2 +
          (dataTemp$IncorrectA1X2)**2 + (dataTemp$IncorrectB1X2)**2 )
578   p1.season.1x2 <- c(p1.season.1x2, round(p1.temp, 5))
579   p2.season.1x2 <- c(p2.season.1x2, round(p2.temp ,5))
580
581   p1.uo.temp <- exp(1/(nrow(dataTemp)) * sum(dataTemp$LogCorrectUO))
582   p2.uo.temp <- 1/(nrow(dataTemp))*sum((1 - dataTemp$CorrectUO)**2 +
          (dataTemp$IncorrectUO)**2)
583   p1.season.uo<- c(p1.season.uo, round(p1.uo.temp, 5))
584   p2.season.uo <- c(p2.season.uo, round(p2.uo.temp, 5))
585
586   dataTempAH <- ensco.ah.results[ensco.ah.results$Season == i, ]
587
588   p1.ah.temp <- exp(1/(nrow(dataTempAH)) * sum(dataTempAH$LogCorrectAH))
589   p2.ah.temp <- 1/(nrow(dataTempAH)) * sum((1 - dataTempAH$CorrectAH)**2 +
          (dataTempAH$IncorrectAH)**2)
590   p1.season.ah <- c(p1.season.ah, round(p1.ah.temp, 5))
591   p2.season.ah <- c(p2.season.ah, round(p2.ah.temp, 5))
592
593   slope.season.1x2 <- c(slope.season.1x2, modelTemp.1x2$coefficients[2]) #Coefficients[2]
          is the gradient
594   slope.season.uo <- c(slope.season.uo, modelTemp.uo$coefficients[2])
595   slope.season.ah <- c(slope.season.ah, modelTemp.ah$coefficients[2])
596 }
597 rmse.season <- matrix( c(rmse.season.1x2, rmse.season.uo, rmse.season.ah), ncol=3,
        byrow=F, dimnames = list(seasons, c("1x2", "UO", "AH")))
598 rsqu.season <- matrix( c(rsqu.season.1x2, rsqu.season.uo, rsqu.season.ah), ncol=3,
        byrow=F, dimnames = list(seasons, c("1x2", "UO", "AH")))
599 p.values.season <- matrix(c(p1.season.1x2, p2.season.1x2, p1.season.uo, p2.season.uo,
        p1.season.ah, p2.season.ah), ncol=15, byrow=T, dimnames = list(c("P1 1X2", "P2 1X2",
        "P1 UO", "P2 UO", "P1 AH", "P2 AH"), seasons))
600 slope.season <- matrix(c(slope.season.1x2, slope.season.uo, slope.season.ah), ncol = 3,
        byrow = F, dimnames = list(seasons, c("1X2", "UO", "AH")))
601
602 #To present these in the text, we use the following two matrices (lots of values => split
        over 2)
603 accuracyvaluematrix.byseason1 <- round(matrix(c(rsqu.season[,1], rmse.season[,1],
```

```
    p1.season, p2.season, slope.season[,1], rsqu.season[,2], rmse.season[,2],
    p1.season.uo, p2.season.uo, slope.season[,2]), nrow = 15, byrow = F), 4)
rownames(accuracyvaluematrix.byseason1) <- seasons
accuracyvaluematrix.byseason2 <- round(matrix(c(rsqu.season[,3], rmse.season[,3],
    p1.season.ah, p2.season.ah, slope.season[,3]), nrow = 15, byrow = F), 4)
rownames(accuracyvaluematrix.byseason2) <- seasons



#Plotting these values -- We make 15 plots and arrange onto one figure after:
rsqu.season.plot.1x2 <- ggplot(NULL, aes(y=rsqu.season.1x2, x=c(2005:2019))) +
    geom_jitter(color = "violetred1") + theme_light() + labs(x = 'Year', y = 'R2', title =
    'R2, 1X2') + geom_smooth(method = 'lm', se = F, color = "violetred1")

rmse.season.plot.1x2 <- ggplot(NULL, aes(y=rmse.season.1x2, x=c(2005:2019))) +
    geom_jitter(color = "violetred1") + theme_light() + labs(x = 'Year', y = 'RMSE', title
    = 'RMSE, 1X2') + geom_smooth(method = 'lm', se = F, color = "violetred1")

p1.season.plot.1x2 <- ggplot(NULL, aes(y=p1.season.1x2, x=c(2005:2019))) +
    geom_jitter(color = "violetred1") + theme_light() + labs(x = 'Year', y = 'P1', title =
    'P1, 1X2') + geom_smooth(method = 'lm', se = F, color = "violetred1")

p2.season.plot.1x2 <- ggplot(NULL, aes(y=p2.season.1x2, x=c(2005:2019))) +
    geom_jitter(color = "violetred1") + theme_light() + labs(x = 'Year', y = 'P2', title =
    'P2, 1X2') + geom_smooth(method = 'lm', se = F, color = "violetred1")

slope.season.plot.1x2 <- ggplot(NULL, aes(y=slope.season.1x2, x=c(2005:2019))) +
    geom_jitter(color = "violetred1") + theme_light() + labs(x = 'Year', y = 'Slope',
    title = 'Slope, 1X2') + geom_smooth(method = 'lm', se = F, color = "violetred1") +
    geom_abline(slope = 0, intercept = 1, color = "black")

rsqu.season.plot.uo <- ggplot(NULL, aes(y=rsqu.season.uo, x=c(2005:2019))) +
    geom_jitter(color = "dodgerblue4") + theme_light() + labs(x = 'Year', y = 'R2', title
    = 'R2, UO') + geom_smooth(method = 'lm', se = F, color = "dodgerblue4")

rmse.season.plot.uo <- ggplot(NULL, aes(y=rmse.season.uo, x=c(2005:2019))) +
    geom_jitter(color = "dodgerblue4") + theme_light() + labs(x = 'Year', y = 'RMSE',
    title = 'RMSE, UO') + geom_smooth(method = 'lm', se = F, color = "dodgerblue4")

p1.season.plot.uo <- ggplot(NULL, aes(y=p1.season.uo, x=c(2005:2019))) + geom_jitter(color
    = "dodgerblue4") + theme_light() + labs(x = 'Year', y = 'P1', title = 'P1, UO') +
    geom_smooth(method = 'lm', se = F, color = "dodgerblue4")

p2.season.plot.uo <- ggplot(NULL, aes(y=p2.season.uo, x=c(2005:2019))) + geom_jitter(color
    = "dodgerblue4") + theme_light() + labs(x = 'Year', y = 'P2', title = 'P2, UO') +
    geom_smooth(method = 'lm', se = F, color = "dodgerblue4")

slope.season.plot.uo <- ggplot(NULL, aes(y=slope.season.uo, x=c(2005:2019))) +
    geom_jitter(color = "dodgerblue4") + theme_light() + labs(x = 'Year', y = 'Slope',
    title = 'Slope, UO') + geom_smooth(method = 'lm', se = F, color = "dodgerblue4") +
    geom_abline(slope = 0, intercept = 1, color = "black")

```

```
631 rsqu.season.plot.ah <- ggplot(NULL, aes(y=rsqu.season.ah, x=c(2005:2019))) +
        geom_jitter(color = "seagreen4") + theme_light() + labs(x = 'Year', y = 'R2', title =
        'R2, AH') + geom_smooth(method = 'lm', se = F, color = "seagreen4")
632
633 rmse.season.plot.ah <- ggplot(NULL, aes(y=rmse.season.ah, x=c(2005:2019))) +
        geom_jitter(color = "seagreen4") + theme_light() + labs(x = 'Year', y = 'RMSE', title
        = 'RMSE, AH') + geom_smooth(method = 'lm', se = F, color = "seagreen4")
634
635 p1.season.plot.ah <- ggplot(NULL, aes(y=p1.season.ah, x=c(2005:2019))) + geom_jitter(color
        = "seagreen4") + theme_light() + labs(x = 'Year', y = 'P1', title = 'P1, AH') +
        geom_smooth(method = 'lm', se = F, color = "seagreen4")
636
637 p2.season.plot.ah <- ggplot(NULL, aes(y=p2.season.ah, x=c(2005:2019))) + geom_jitter(color
        = "seagreen4") + theme_light() + labs(x = 'Year', y = 'P2', title = 'P2, AH') +
        geom_smooth(method = 'lm', se = F, color = "seagreen4")
638
639 slope.season.plot.ah <- ggplot(NULL, aes(y=slope.season.ah, x=c(2005:2019))) +
        geom_jitter(color = "seagreen4") + theme_light() + labs(x = 'Year', y = 'Slope', title
        = 'Slope, AH') + geom_smooth(method = 'lm', se = F, color = "seagreen4") +
        geom_abline(slope = 0, intercept = 1, color = "black")
640
641 seasontimeplot <- grid.arrange(rsqu.season.plot.1x2, rsqu.season.plot.uo,
        rsqu.season.plot.ah,
642                            rmse.season.plot.1x2, rmse.season.plot.uo, rmse.season.plot.ah,
643                            p1.season.plot.1x2, p1.season.plot.uo, p1.season.plot.ah,
644                            p2.season.plot.1x2, p2.season.plot.uo, p2.season.plot.ah,
645                            slope.season.plot.1x2, slope.season.plot.uo,
                                slope.season.plot.ah,
646                            nrow = 5, top = 'English & Scottish Leagues Accuracy
                                Statistics over Time')
647 ggsave(path="./writeup/images", filename="ensco_20_seasontimeplots.png",
        plot=seasontimeplot, unit="cm", width=20, height=25)
648
649
650 ### OVERROUND ----
651 #Plots
652 or.ot.l <- ggplot(ensco, aes(x=AvgHProb, y=OneXTwoOverround, color = Level)) +
        geom_jitter(alpha = 0.5) + theme_light() + guides(col = guide_legend(ncol = 3)) +
        labs(x = "Consensus P(Home Win)", y = "Sum of Probabilities (1X2)", title = "Consensus
        P(Home Win) v. Bookmaker Commission, by Level", caption = "English/Scottish Leagues,
        2005-20") + coord_cartesian(ylim = c(1, 1.3))
653
654 or.ot.s <- ggplot(ensco, aes(x=AvgHProb, y=OneXTwoOverround, color = Season)) +
        geom_jitter(alpha = 0.5) + theme_light() + guides(col = guide_legend(ncol = 3)) +
        labs(x = "Consensus P(Home Win)", y = "Sum of Probabilities (1X2)", title = "Consensus
        P(Home Win) v. Bookmaker Commission, by Season", caption = "English/Scottish Leagues,
        2005-20") + coord_cartesian(ylim = c(1, 1.3))
655
656 or.uo.l <- ggplot(ensco,aes(x=Over2.5Prob,y=UnderOverOverround,color=Level)) +
        geom_jitter(alpha = 0.5) + theme_light() + guides(col = guide_legend(ncol = 3)) +
        labs(x = "Consensus P(Over 2.5 Goals)", y="Sum of Probabilities (UO)", title =
        "Consensus P(Over 2.5 Goals) v. Bookmaker Commission, by Level", caption =
```

```
    "English/Scottish Leagues, 2005-20") + coord_cartesian(ylim = c(1, 1.3))
657
658 or.uo.s <- ggplot(ensco,aes(x=Over2.5Prob,y=UnderOverOverround,color=Season)) +
    geom_jitter(alpha = 0.5) + theme_light() + guides(col = guide_legend(ncol = 3)) +
    labs(x = "Consensus P(Over 2.5 Goals)", y="Sum of Probabilities (UO)", title =
    "Consensus P(Over 2.5 Goals) v. Bookmaker Commission, by Season", caption =
    "English/Scottish Leagues, 2005-20") + coord_cartesian(ylim = c(1, 1.3))
659
660 or.ah.l <- ggplot(ensco, aes(x=AH.HProb, y=AHOverround, color = Level)) +
    geom_jitter(alpha = 0.5) + theme_light() + guides(col = guide_legend(ncol = 3)) +
    labs(x = "Consensus P(AH Home Win)", y = "Sum of Probabilities (AH)", title =
    "Consensus P(Home Win, AH) v. Bookmaker Commission, by Level", caption =
    "English/Scottish Leagues, 2005-20") + coord_cartesian(ylim = c(1, 1.3))
661
662 or.ah.s <- ggplot(ensco, aes(x=AH.HProb, y=AHOverround, color = Season)) +
    geom_jitter(alpha = 0.5) + theme_light() + guides(col = guide_legend(ncol = 3)) +
    labs(x = "Consensus P(AH Home Win)", y = "Sum of Probabilities (AH)", title =
    "Consensus P(Home Win, AH) v. Bookmaker Commission, by Season", caption =
    "English/Scottish Leagues, 2005-20") + coord_cartesian(ylim = c(1, 1.3))
663
664 ggsave(path = "./writeup/images", filename = "ensco_14a_overround_ot_l.png", plot=or.ot.l,
    unit="cm", width=20, height=10)
665 ggsave(path = "./writeup/images", filename = "ensco_14b_overround_ot_s.png", plot=or.ot.s,
    unit="cm", width=20, height=10)
666 ggsave(path = "./writeup/images", filename = "ensco_14c_overround_uo_l.png", plot=or.uo.l,
    unit="cm", width=20, height=10)
667 ggsave(path = "./writeup/images", filename = "ensco_14d_overround_uo_s.png", plot=or.uo.s,
    unit="cm", width=20, height=10)
668 ggsave(path = "./writeup/images", filename = "ensco_14e_overround_ah_l.png", plot=or.ah.l,
    unit="cm", width=20, height=10)
669 ggsave(path = "./writeup/images", filename = "ensco_14f_overround_ah_s.png", plot=or.ah.s,
    unit="cm", width=20, height=10)
670
671
672 #Overround Calcs
673 #Overall, across all leagues/season:
674 mean(ensco$OneXTwoOverround)
675 mean(ensco$UnderOverOverround)
676 mean(ensco$AHOverround)
677
678 #By Level
679 for (i in levels){
680   print(paste("Level",i,"-------------------------------------------"))
681   print(paste("1X2 Overround: ",mean(ensco[ensco$Level==i,]$OneXTwoOverround)))
682   print(paste("UO Overround: ",mean(ensco[ensco$Level==i,]$UnderOverOverround)))
683   print(paste("AH Overround: ",mean(ensco[ensco$Level==i,]$AHOverround)))
684 }
685 #Select seasons (05/06, 12/13 and 19/20) (equally spaced)
686 for (j in c("0506", "1213", "1920")){
687   print(paste("Season",j,"-------------------------------------------"))
688   print(paste("1X2 Overround: ",mean(ensco[ensco$Season==j,]$OneXTwoOverround)))
689   print(paste("UO Overround: ",mean(ensco[ensco$Season==j,]$UnderOverOverround)))
```

```
690   print(paste("AH Overround: ",mean(ensco[ensco$Season==j,]$AHOverround)))
691 }
692
693 #- End -
```

# Appendix D

# Chapter 4 Code

```
#### (3) USING OUR RESULTS FOR A BETTING ALGORITHM
library(ggplot2); library(scales); library(e1071)

## CLEANING AND OBTAINING DATA ----
# First, we need to trim the datasets to just the columns we need and obtain the Asian
    Handicap odds for elite leagues from F-D.co.uk

matchesTemp <- NULL; matches <- NULL
seasons <- c("0506", "0607", "0708", "0809", "0910", "1011", "1112", "1213", "1314",
    "1415", "1516", "1617", "1718", "1819", "1920")
divisions <- c("E0", "E1", "E2", "E3", "EC", "SC0", "SC1", "SC2", "SC3", "D1", "SP1",
    "F1", "I1", "P1")

for (i in seasons){
  for (j in divisions){
    matchesTemp <- read.csv(paste0("https://www.football-data.co.uk/mmz4281/", i, "/", j,
        ".csv"), fileEncoding="latin1")
    #The above line will download and read the .csv file in one go.
    matchesTemp$Season <- with(matchesTemp,i)
    matchesTemp$Div <- with(matchesTemp, j)
    if (i=="1920"){
      matchesTemp$BbAvH <- matchesTemp$AvgH
      matchesTemp$BbAvA <- matchesTemp$AvgA
      matchesTemp$BbAvD <- matchesTemp$AvgD
      matchesTemp$BbAvAHH <- matchesTemp$AvgAHH
      matchesTemp$BbAvAHA <- matchesTemp$AvgAHA
      matchesTemp$BbAHh <- matchesTemp$AHh}
    else{}
    matchesTemp$HomeHandicap <- matchesTemp$BbAHh
    matchesTemp <- matchesTemp[,c("Div", "Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG",
        "FTR", "BbAvH", "BbAvD", "BbAvA", "HomeHandicap", "BbAvAHH", "BbAvAHA", "Season")]
    matches <- rbind(matches, matchesTemp)
  }
}
matches <- na.omit(matches)
```

```
31
32 # Finding underlying probabilities of each event:-
33 matches$OT.HProb.PN <- with(matches, 1/(BbAvH))
34 matches$OT.DProb.PN <- with(matches, 1/(BbAvD))
35 matches$OT.AProb.PN <- with(matches, 1/(BbAvA))
36 matches$AH.HProb.PN <- with(matches, 1/(BbAvAHH))
37 matches$AH.AProb.PN <- with(matches, 1/(BbAvAHA))
38
39 # Finding consensus probabilities of each event:-
40 matches$OT.HProb <- with(matches, round(OT.HProb.PN / (OT.HProb.PN + OT.DProb.PN +
       OT.AProb.PN), 4))
41 matches$OT.AProb <- with(matches, round(OT.AProb.PN / (OT.HProb.PN + OT.DProb.PN +
       OT.AProb.PN), 4))
42 matches$AH.HProb <- with(matches,round(AH.HProb.PN / (AH.HProb.PN + AH.AProb.PN), 4))
43 matches$AH.AProb <- with(matches,round(AH.AProb.PN / (AH.HProb.PN + AH.AProb.PN), 4))
44
45 N = nrow(matches)
46
47 matches$FTHG.ah <- with(matches, rep(0,N))
48 for (m in 1:N){matches$FTHG.ah[m] <- matches$FTHG[m] + matches$HomeHandicap[m]}
49 matches$ah.gap <- with(matches, FTHG.ah - FTAG); matches$ah.res <- NULL
50 for (n in 1:N){
51   if (matches$ah.gap[n]<(-0.25)){matches$ah.res[n] <- "aw"}
52   else if (matches$ah.gap[n]==(-0.25)){matches$ah.res[n] <- "hfaw"}
53   else if (matches$ah.gap[n]==0){matches$ah.res[n] <- "vo"}
54   else if (matches$ah.gap[n]==0.25){matches$ah.res[n] <- "hfhm"}
55   else if (matches$ah.gap[n]>0.25){matches$ah.res[n] <- "hm"}
56   else{}
57 }
58
59
60 ## PLACING BETS ----
61 # To assess our algorithm, we assume we use the first year (05/06) to obtain data (that
       is, the means and std devs) and place bets on any game after.
62 matches$OTHomeBet <- with(matches, 0); matches$OTAwayBet <- with(matches, 0)
63 matches$AHHomeBet <- with(matches, 0); matches$AHAwayBet <- with(matches, 0)
64
65 #Initial Bounds :-
66 matches0506 <- matches[matches$Season == '0506',]
67 mu.oth <- mean(matches0506$OT.HProb); sd.oth <- sd(matches0506$OT.HProb)
68 mu.ota <- mean(matches0506$OT.AProb); sd.ota <- sd(matches0506$OT.AProb)
69 mu.ahh <- mean(matches0506$AH.HProb); sd.ahh <- sd(matches0506$AH.HProb)
70 mu.aha <- mean(matches0506$AH.AProb); sd.aha <- sd(matches0506$AH.AProb)
71 n <- nrow(matches[matches$Season == "0506",]); N <- nrow(matches)
72 #Placing Bets:-
73 for (i in n:N){
74   #Update the mean and std dev's with our new information
75   mu.oth <- mean(matches$OT.HProb[1:i]); sd.oth <- sd(matches$OT.HProb[1:i])
76   mu.ota <- mean(matches$OT.AProb[1:i]); sd.ota <- sd(matches$OT.AProb[1:i])
77   mu.ahh <- mean(matches$AH.HProb[1:i]); sd.ahh <- sd(matches$AH.HProb[1:i])
78   mu.aha <- mean(matches$AH.AProb[1:i]); sd.aha <- sd(matches$AH.AProb[1:i])
79   #Do we bet on Home Win (1X2)?
```

```
  if (matches$OT.HProb[i] > mu.oth + 0.5*sd.oth){
    if (matches$OT.HProb[i] <= mu.oth + sd.oth){matches$OTHomeBet[i] <- 1}
    else if (matches$OT.HProb[i] > mu.oth + sd.oth & matches$OT.HProb[i] <= mu.oth +
        1.5*sd.oth){matches$OTHomeBet[i] <- 2}
    else {matches$OTHomeBet[i] <- 3}}
  else {matches$OTHomeBet[i] <- 0}
  #Do we bet on Away Win (1X2)?
  if (matches$OT.AProb[i] > mu.ota + 0.5*sd.ota){
    if (matches$OT.AProb[i] <= mu.ota + sd.ota){matches$OTAwayBet[i] <- 1}
    else if (matches$OT.AProb[i] > mu.ota + sd.ota & matches$OT.AProb[i] <= mu.ota +
        1.5*sd.ota){matches$OTAwayBet[i] <- 2}
    else {matches$OTAwayBet[i] <- 3}}
  else {matches$OTAwayBet[i] <- 0}
  #Do we bet on Home Win (AH)?
  if (matches$AH.HProb[i] > mu.ahh + 0.5*sd.ahh){
    if (matches$AH.HProb[i] <= mu.ahh + sd.ahh){matches$AHHomeBet[i] <- 1}
    else if (matches$AH.HProb[i] > mu.ahh + sd.ahh & matches$AH.HProb[i] <= mu.ahh +
        1.5*sd.ahh){matches$AHHomeBet[i] <- 2}
    else {matches$AHHomeBet[i] <- 3}}
  else {matches$AHHomeBet[i] <- 0}
  #Do we bet on Away Win (AH)?
  if (matches$AH.AProb[i] > mu.aha + 0.5*sd.aha){
    if (matches$AH.AProb[i] <= mu.aha + sd.aha){matches$AHAwayBet[i] <- 1}
    else if (matches$AH.AProb[i] > mu.aha + sd.aha & matches$AH.AProb[i] <= mu.aha +
        1.5*sd.aha){matches$AHAwayBet[i] <- 2}
    else {matches$AHAwayBet[i] <- 3}}
  else {matches$AHAwayBet[i] <- 0}
}

## FINDING RESULTS ----
matches$OTHReturns <- with(matches, 0); matches$OTAReturns <- with(matches, 0)
matches$AHHReturns <- with(matches, 0); matches$AHAReturns <- with(matches, 0)

for (i in n:N){
  if (matches$FTR[i]=="H"){
    matches$OTHReturns[i] <- (matches$BbAvH[i] - 1) * matches$OTHomeBet[i]
    matches$OTAReturns[i] <- -matches$OTAwayBet[i]}
  else if (matches$FTR[i]=="A"){
    matches$OTAReturns[i] <- (matches$BbAvA[i] - 1) * matches$OTAwayBet[i]
    matches$OTHReturns[i] <- -matches$OTHomeBet[i]}
  else {
    matches$OTHReturns[i] <- -matches$OTHomeBet[i]
    matches$OTAReturns[i] <- -matches$OTAwayBet[i]}
}

for (i in n:N){
  if (matches$ah.res[i]=="aw"){
    matches$AHAReturns[i] <- (matches$BbAvAHA[i] - 1) * matches$AHAwayBet[i]
    matches$AHHReturns[i] <- -matches$AHHomeBet[i]
  }
  else if (matches$ah.res[i]=="hfaw"){
    matches$AHAReturns[i] <- (matches$BbAvAHA[i] - 1) * 0.5 * matches$AHAwayBet[i] - 0.5 *
```

```
        matches$AHAwayBet[i]
    matches$AHHReturns[i] <- -matches$AHHomeBet[i]
  }
  else if (matches$ah.res[i]=="hm"){
    matches$AHHReturns[i] <- (matches$BbAvAHH[i] - 1) * matches$AHHomeBet[i]
    matches$AHAReturns[i] <- -matches$AHAwayBet[i]
  }
  else if (matches$ah.res[i]=="hfhm"){
    matches$AHHReturns[i] <- (matches$BbAvAHH[i] - 1) * 0.5 * matches$AHHomeBet[i] - 0.5 *
        matches$AHHomeBet[i]
    matches$AHAReturns[i] <- -matches$AHAwayBet[i]
  }
  else{ #i.e. Void
    matches$AHHReturns[i] <- 0
    matches$AHAReturns[i] <- 0
  }
}

#Cumulative Returns (for our plot):-
matches$C.OTHReturns <- with(matches, 0)
matches$C.OTAReturns <- with(matches, 0)
matches$C.AHHReturns <- with(matches, 0)
matches$C.AHAReturns <- with(matches, 0)
matches$C.Returns <- with(matches, 0)

matches$C.OTHReturns[1] <- matches$OTHReturns[1]
matches$C.OTAReturns[1] <- matches$OTAReturns[1]
matches$C.AHHReturns[1] <- matches$AHHReturns[1]
matches$C.AHAReturns[1] <- matches$AHAReturns[1]
matches$C.Returns[1] <- matches$OTHReturns[1] + matches$OTAReturns[1] +
    matches$AHHReturns[1] + matches$AHAReturns[1]

for (i in 2:N){matches$C.Returns[i] <- matches$C.Returns[i-1] + matches$OTHReturns[i] +
    matches$OTAReturns[i] + matches$AHHReturns[i] + matches$AHAReturns[i]}
for (i in 2:N){matches$C.OTHReturns[i] <- matches$C.OTHReturns[i-1] +
    matches$OTHReturns[i]}
for (i in 2:N){matches$C.OTAReturns[i] <- matches$C.OTAReturns[i-1] +
    matches$OTAReturns[i]}
for (i in 2:N){matches$C.AHHReturns[i] <- matches$C.AHHReturns[i-1] +
    matches$AHHReturns[i]}
for (i in 2:N){matches$C.AHAReturns[i] <- matches$C.AHAReturns[i-1] +
    matches$AHAReturns[i]}

matches$BetIndex <- with(matches,0)
for (i in n:N){matches$BetIndex[i] <- (i-(n-1))}

## PLOTS ----
cr1 <- ggplot(matches,aes(x=BetIndex,y=C.OTHReturns,color="1X2 H Returns")) +
  geom_line() +
  geom_line(aes(x=BetIndex, y=C.OTAReturns, color="1X2 A Returns")) +
  geom_line(aes(x=BetIndex, y=C.AHHReturns, color="AH H Returns")) +
  geom_line(aes(x=BetIndex, y=C.AHAReturns, color="AH A Returns")) +
```

```
  #geom_line(aes(x=BetIndex, y=C.Returns, color="Total Returns"), linetype="dotted") +
  theme_light() + labs(title="Cumulative Winnings using our Algorithm", caption="All matches, 2006-20", x="Match Index", y="Returns (unit)") +
  scale_color_manual(name="Bet Type", values=c("1X2 H Returns" = "blue", "1X2 A Returns" = "coral", "AH H Returns" = "green", "AH A Returns" = "purple", "Total Returns" = "black")) + geom_hline(yintercept=0, color="black", linetype="dotted", alpha=.5)

cr2 <- ggplot(matches, aes(x=BetIndex, y=C.OTHReturns, color="1X2 H Returns")) +
  geom_line() +
  geom_line(aes(x=BetIndex, y=C.OTAReturns, color="1X2 A Returns")) +
  geom_line(aes(x=BetIndex, y=C.AHHReturns, color="AH H Returns")) +
  geom_line(aes(x=BetIndex, y=C.AHAReturns, color="AH A Returns")) +
  geom_line(aes(x=BetIndex, y=C.Returns, color="Total Returns"), linetype="twodash") +
  scale_color_manual(name="Bet Type", values=c("1X2 H Returns" = "blue", "1X2 A Returns" = "coral", "AH H Returns" = "green", "AH A Returns" = "purple", "Total Returns" = "black")) +
  theme_light() + coord_cartesian(xlim=c(0, 100), ylim=c(-25,25)) +
      geom_hline(yintercept=0, color="black", linetype="dotted", alpha=.5) +
      labs(title="Cumulative Winnings (First 100 Games)\n using our Algorithm",
      caption="All matches from 2006-20", x="Match Index", y="Returns (unit)")

ggsave(path = "./writeup/images", filename = "model_01.png", plot=cr1, unit="cm", width=15, height=15)
ggsave(path = "./writeup/images", filename = "model_02.png", plot=cr2, unit="cm", width=15, height=10)

## ACCURACY OF THE ALGORITHM ----
n.OTH <- nrow(matches[matches$OTHomeBet > 0,])
n.OTA <- nrow(matches[matches$OTAwayBet > 0,])
n.AHH <- nrow(matches[matches$AHHomeBet > 0,])
n.AHA <- nrow(matches[matches$AHAwayBet > 0,])
n.BetsPlaced <- n.OTH + n.OTA + n.AHH + n.AHA
n.MatchesBet <- nrow(matches[which(matches$OTHomeBet > 0 | matches$OTAwayBet > 0 | matches$AHHomeBet > 0 | matches$AHAwayBet > 0),])
n.BetsMatrix <- matrix(round(c(n.OTH, n.OTA, n.AHH, n.AHA, n.BetsPlaced), 5), ncol = 5)

accuracy.ot.h <- nrow(matches[matches$OTHReturns > 0,]) / n.OTH * 100
accuracy.ot.a <- nrow(matches[matches$OTAReturns > 0,]) / n.OTA * 100
accuracy.ah.h <- nrow(matches[matches$AHHReturns > 0,]) / n.AHH * 100
accuracy.ah.a <- nrow(matches[matches$AHAReturns > 0,]) / n.AHA * 100
accuracy.overall <- 100 * (nrow(matches[matches$OTHReturns > 0,]) + nrow(matches[matches$OTAReturns > 0,]) + nrow(matches[matches$AHHReturns > 0,]) + nrow(matches[matches$AHAReturns > 0,])) / n.BetsPlaced

accuracy <- matrix(c(accuracy.ot.h, accuracy.ot.a, accuracy.ah.h, accuracy.ah.a, accuracy.overall), ncol = 5)

total.winnings <- sum(matches$OTHReturns) + sum(matches$OTAReturns) + sum(matches$AHHReturns) + sum(matches$AHAReturns)

bet.winnings <- matrix(c(sum(matches$OTHReturns), sum(matches$OTAReturns), sum(matches$AHHReturns), sum(matches$AHAReturns), total.winnings), ncol = 5)
```

```
208
209 bet.analysis <- matrix(c(n.BetsMatrix, bet.winnings, accuracy), nrow = 5, byrow = F,
        dimnames = list(c('1x2 H', '1x2 A', 'AH H', 'AH A', 'Overall'), c('Bets
        Placed','Winnings', 'Accuracy (%)')))
210 bet.analysis
211
212 ## IGNORING THE 'LOW' PERFORMING LEAGUES (ALTERNATIVE METHOD) ----
213 #From our analysis, whilst all leagues performed well, we noticed the German and French,
        and English/Scottish Level 2 Leagues.
214
215 #First, we copy the bets we placed earlier into new columns:
216 matches$OTHomeBet.alt <- with(matches, OTHomeBet)
217 matches$OTAwayBet.alt <- with(matches, OTAwayBet)
218 matches$AHHomeBet.alt <- with(matches, AHHomeBet)
219 matches$AHAwayBet.alt <- with(matches, AHAwayBet)
220
221 #Removing bets placed in France, Germany and Level 2, Eng/Sco:-
222 for (i in 1:N){
223   if (matches$Div[i] %in% c("D1", "F1", "E2", "E3", "SC1")){
224     matches$OTHomeBet.alt[i] <- 0
225     matches$OTAwayBet.alt[i] <- 0
226     matches$AHHomeBet.alt[i] <- 0
227     matches$AHAwayBet.alt[i] <- 0
228   }
229 }
230
231 #And doing the same with returns:-
232 matches$OTHRet.alt <- with(matches, OTHReturns)
233 matches$OTARet.alt <- with(matches, OTAReturns)
234 matches$AHHRet.alt <- with(matches, AHHReturns)
235 matches$AHARet.alt <- with(matches, AHAReturns)
236 for (i in 1:N){
237   if (matches$Div[i] %in% c("D1", "F1", "E2", "E3", "SC1")){
238     matches$OTHRet.alt[i] <- 0
239     matches$OTARet.alt[i] <- 0
240     matches$AHHRet.alt[i] <- 0
241     matches$AHARet.alt[i] <- 0
242   }
243 }
244 #Cumulative Returns
245 matches$C.OTHRet.alt <- with(matches,0); matches$C.OTARet.alt <- with(matches,0);
        matches$C.AHHRet.alt <- with(matches,0); matches$C.AHARet.alt <- with(matches,0);
        matches$C.Returns.alt <- with(matches, 0)
246
247 matches$C.OTHRet.alt[1] <- matches$OTHRet.alt[1]
248 matches$C.OTARet.alt[1] <- matches$OTARet.alt[1]
249 matches$C.AHHRet.alt[1] <- matches$AHHRet.alt[1]
250 matches$C.AHARet.alt[1] <- matches$AHARet.alt[1]
251 matches$C.Returns.alt[1] <- matches$OTHRet.alt[1] + matches$OTARet.alt[1] +
        matches$AHHRet.alt[1] + matches$AHARet.alt[1]
252
253 for (i in 2:N){matches$C.Returns.alt[i] <- matches$C.Returns.alt[i-1] +
```

```
    matches$OTHRet.alt[i] + matches$OTARet.alt[i] + matches$AHHRet.alt[i] +
    matches$AHARet.alt[i]}
for (i in 2:N){matches$C.OTHRet.alt[i] <- matches$C.OTHRet.alt[i-1] +
    matches$OTHRet.alt[i]}
for (i in 2:N){matches$C.OTARet.alt[i] <- matches$C.OTARet.alt[i-1] +
    matches$OTARet.alt[i]}
for (i in 2:N){matches$C.AHHRet.alt[i] <- matches$C.AHHRet.alt[i-1] +
    matches$AHHRet.alt[i]}
for (i in 2:N){matches$C.AHARet.alt[i] <- matches$C.AHARet.alt[i-1] +
    matches$AHARet.alt[i]}

#Accuracy
n.OTH.alt <- nrow(matches[matches$OTHomeBet.alt > 0,])
n.OTA.alt <- nrow(matches[matches$OTAwayBet.alt > 0,])
n.AHH.alt <- nrow(matches[matches$AHHomeBet.alt > 0,])
n.AHA.alt <- nrow(matches[matches$AHAwayBet.alt > 0,])
n.BetsP.alt <- n.OTH.alt + n.OTA.alt + n.AHH.alt + n.AHA.alt
n.BetsMatrix.alt <- matrix(round(c(n.OTH.alt, n.OTA.alt, n.AHH.alt, n.AHA.alt,
    n.BetsP.alt), 5), ncol = 5)

accalt.ot.h <- nrow(matches[matches$OTHRet.alt > 0,]) / n.OTH.alt * 100
accalt.ot.a <- nrow(matches[matches$OTARet.alt > 0,]) / n.OTA.alt * 100
accalt.ah.h <- nrow(matches[matches$AHHRet.alt > 0,]) / n.AHH.alt * 100
accalt.ah.a <- nrow(matches[matches$AHARet.alt > 0,]) / n.AHA.alt * 100

accalt.ovr <- 100 * (nrow(matches[matches$OTHRet.alt > 0,]) +
    nrow(matches[matches$OTARet.alt > 0,]) + nrow(matches[matches$AHHRet.alt > 0,]) +
    nrow(matches[matches$AHARet.alt > 0,])) / n.BetsP.alt
accalt <- matrix(c(accalt.ot.h, accalt.ot.a, accalt.ah.h, accalt.ah.a, accalt.ovr), ncol =
    5)
total.wins.alt <- sum(matches$OTHRet.alt) + sum(matches$OTARet.alt) +
    sum(matches$AHHRet.alt) + sum(matches$AHARet.alt)
bet.wins.alt <- matrix(c(sum(matches$OTHRet.alt), sum(matches$OTARet.alt),
    sum(matches$AHHRet.alt), sum(matches$AHARet.alt), total.wins.alt), ncol = 5)
bet.analysis.alt <- matrix(c(n.BetsMatrix.alt, bet.wins.alt, accalt), nrow = 5, byrow = F,
    dimnames = list(c('1x2 H', '1x2 A', 'AH H', 'AH A', 'Overall'), c('Bets
    Placed','Winnings', 'Accuracy (%)')))

bet.analysis.alt

## COMPARISON OF ACCURACY AGAINST RANDOMLY PLACED BETS ----
#To compare our algorithm against randomly placed bets, we will choose a random subset of
    N matches and find the winnings.

nRunRBS <- 10

N <- nrow(matches)
for (i in 1:nRunRBS){
  #Reset probabilities for 1X2 Home
  p1 <- nrow(matches[matches$OTHomeBet == 1,]) / nrow(matches[(n+1):N,])
  p2 <- nrow(matches[matches$OTHomeBet == 2,]) / nrow(matches[(n+1):N,])
  p3 <- nrow(matches[matches$OTHomeBet == 3,]) / nrow(matches[(n+1):N,])
```

```
291 p0 <- 1 - (p1 + p2 + p3)
292
293 matches$rand.Bet.OTH <- with(matches, 0)
294 matches$rand.Ret.OTH <- with(matches, 0)
295 matches$rand.Bet.OTH <- with(matches, rand.Bet.OTH + rdiscrete(n = nrow(matches), values
        = 0:3, probs=c(p0, p1, p2, p3)))
296
297 for(i in n:N){
298   if (matches$FTR[i] == "H"){matches$rand.Ret.OTH[i] <- (matches$BbAvH[i]-1) *
          matches$rand.Bet.OTH[i]}
299   else{matches$rand.Ret.OTH[i] <- -matches$rand.Bet.OTH[i]}}
300
301 #Reset probabilities for 1X2 Away
302 p1 <- nrow(matches[matches$OTAwayBet == 1,]) / nrow(matches[(n+1):N,])
303 p2 <- nrow(matches[matches$OTAwayBet == 2,]) / nrow(matches[(n+1):N,])
304 p3 <- nrow(matches[matches$OTAwayBet == 3,]) / nrow(matches[(n+1):N,])
305 p0 <- 1 - (p1 + p2 + p3)
306
307 matches$rand.Bet.OTA <- with(matches, 0)
308 matches$rand.Ret.OTA <- with(matches, 0)
309 matches$rand.Bet.OTA <- with(matches, rand.Bet.OTA + rdiscrete(n = nrow(matches), values
        = 0:3, probs=c(p0, p1, p2, p3)))
310
311 for (i in n:N){
312   if (matches$FTR[i] == "A"){matches$rand.Ret.OTA[i] <- (matches$BbAvA[i]-1) *
          matches$rand.Bet.OTA[i]}
313   else{matches$rand.Ret.OTA[i] <- -matches$rand.Bet.OTA[i]}}
314
315 #Reset probabilities for AH Home
316 p1 <- nrow(matches[matches$AHHomeBet == 1,]) / nrow(matches[(n+1):N,])
317 p2 <- nrow(matches[matches$AHHomeBet == 2,]) / nrow(matches[(n+1):N,])
318 p3 <- nrow(matches[matches$AHHomeBet == 3,]) / nrow(matches[(n+1):N,])
319 p0 <- 1 - (p1 + p2 + p3)
320
321 matches$rand.Bet.AHH <- with(matches, 0)
322 matches$rand.Ret.AHH <- with(matches, 0)
323 matches$rand.Bet.AHH <- with(matches, rand.Bet.AHH + rdiscrete(n = nrow(matches), values
        = 0:3, probs=c(p0, p1, p2, p3)))
324 for (i in n:N){
325   if (matches$ah.res[i] == "hm"){matches$rand.Ret.AHH[i] <- (matches$BbAvAHH[i]-1) *
          matches$rand.Bet.AHH[i]}
326   else if (matches$ah.res[i] == "hfhm"){matches$rand.Ret.AHH[i] <- (matches$BbAvAHH[i]-1)
          * 0.5 * matches$rand.Bet.AHH[i] - (0.5 * matches$rand.Bet.AHH[i])}
327   else{matches$rand.Ret.AHH[i] <- -matches$rand.Bet.AHH[i]}}
328
329 #Reset probabilities for AH Away
330 p1 <- nrow(matches[matches$AHAwayBet == 1,]) / nrow(matches[(n+1):N,])
331 p2 <- nrow(matches[matches$AHAwayBet == 2,]) / nrow(matches[(n+1):N,])
332 p3 <- nrow(matches[matches$AHAwayBet == 3,]) / nrow(matches[(n+1):N,])
333 p0 <- 1 - (p1 + p2 + p3)
334
335 matches$rand.Bet.AHA <- with(matches, 0)
```

```
336   matches$rand.Ret.AHA <- with(matches, 0)
337   matches$rand.Bet.AHA <- with(matches, rand.Bet.AHA + rdiscrete(n = nrow(matches), values
          = 0:3, probs=c(p0, p1, p2, p3)))
338   for (i in n:N){
339     if (matches$ah.res[i] == "aw"){matches$rand.Ret.AHA[i] <- (matches$BbAvAHA[i]-1) *
            matches$rand.Bet.AHA[i]}
340     else if (matches$ah.res[i] == "hfaw"){matches$rand.Ret.AHA[i] <- (matches$BbAvAHA[i]-1)
            * 0.5 * matches$rand.Bet.AHA[i] - (0.5 * matches$rand.Bet.AHA[i])}
341     else{matches$rand.Ret.AHA[i] <- -matches$rand.Bet.AHA[i]}}
342
343   #Creating our analysis matrix:
344   #No. of bets
345   #We add the caveat of BetIndex > 0 to avoid betting on the 05/06 season
346   n.Ran.OTH <- nrow(matches[matches$rand.Bet.OTH > 0 & matches$BetIndex > 0,])
347   n.Ran.OTA <- nrow(matches[matches$rand.Bet.OTA > 0 & matches$BetIndex > 0,])
348   n.Ran.AHH <- nrow(matches[matches$rand.Bet.AHH > 0 & matches$BetIndex > 0,])
349   n.Ran.AHA <- nrow(matches[matches$rand.Bet.AHA > 0 & matches$BetIndex > 0,])
350   n.RanBets <- n.Ran.OTH + n.Ran.OTA + n.Ran.AHH + n.Ran.AHA
351   n.RanMatx <- matrix(c(n.Ran.OTH, n.Ran.OTA, n.Ran.AHH, n.Ran.AHA, n.RanBets), ncol = 5)
352
353   stake.OTH <- sum(matches$rand.Bet.OTH)
354   stake.OTA <- sum(matches$rand.Bet.OTA)
355   stake.AHH <- sum(matches$rand.Bet.AHH)
356   stake.AHA <- sum(matches$rand.Bet.AHA)
357   stake.ovr <- stake.OTH + stake.OTA + stake.AHH + stake.AHA
358   stakeMatx <- matrix(c(stake.OTH, stake.OTA, stake.AHH, stake.AHA, stake.ovr), ncol = 5)
359
360   #Accuracy percentage
361   acc.Ran.OTH <- nrow(matches[matches$rand.Ret.OTH > 0,]) / n.Ran.OTH * 100
362   acc.Ran.OTA <- nrow(matches[matches$rand.Ret.OTA > 0,]) / n.Ran.OTA * 100
363   acc.Ran.AHH <- nrow(matches[matches$rand.Ret.AHH > 0,]) / n.Ran.AHH * 100
364   acc.Ran.AHA <- nrow(matches[matches$rand.Ret.AHA > 0,]) / n.Ran.AHA * 100
365   acc.Ran.ovr <- 100 * (nrow(matches[matches$rand.Ret.OTH > 0,]) +
          nrow(matches[matches$rand.Ret.OTA > 0,]) + nrow(matches[matches$rand.Ret.AHH > 0,])
          + nrow(matches[matches$rand.Ret.AHA > 0,])) / n.RanBets
366   acc.Ran <- matrix(c(acc.Ran.OTH, acc.Ran.OTA, acc.Ran.AHH, acc.Ran.AHA, acc.Ran.ovr),
          ncol = 5)
367
368   #Winnings
369   ran.winnings <- sum(matches$rand.Ret.OTH) + sum(matches$rand.Ret.OTA) +
          sum(matches$rand.Ret.AHH) + sum(matches$rand.Ret.AHA)
370   ran.winningsMtx <- matrix(c(sum(matches$rand.Ret.OTH), sum(matches$rand.Ret.OTA),
          sum(matches$rand.Ret.AHH), sum(matches$rand.Ret.AHA), ran.winnings), ncol = 5)
371
372   bet.analysis.random <- matrix(c(n.RanMatx, stakeMatx, ran.winningsMtx, acc.Ran), nrow =
          5, byrow = F, dimnames = list(c('1x2 H', '1x2 A', 'AH H', 'AH A', 'Overall'),
          c('Bets Placed', 'Stake', 'Winnings', 'Accuracy (%)')))
373   print(bet.analysis.random)
374 }
375
376 ## PLOTTING OF RANDOM BET STRATEGY ----
377
```

```
#Run par(mfrow=c(2,2)) if you want all four plots in one graph. Else, it is recommended to run each plot separately.
par(mfrow = c(2,2)); nRuns <- 30; alpha0 <- 1/nRuns
#nRuns is the No. runs you wish to have. More = slower.

#1X2 Home
plot(matches$BetIndex, matches$C.OTHReturns, col = alpha("red"), type = 'l', ylim = c(-3000, 10), xlab = 'Index', ylab = 'Returns', main = '1X2 Home Win')
for (i in 1:nRuns){
  #Reset the random bet, 1X2 Home
  p1 <- nrow(matches[matches$OTHomeBet == 1,]) / nrow(matches[(n+1):N,])
  p2 <- nrow(matches[matches$OTHomeBet == 2,]) / nrow(matches[(n+1):N,])
  p3 <- nrow(matches[matches$OTHomeBet == 3,]) / nrow(matches[(n+1):N,])
  p0 <- 1 - (p1 + p2 + p3)

  matches$rand.Bet.OTH <- with(matches, 0)
  matches$rand.Ret.OTH <- with(matches, 0)
  matches$rand.Bet.OTH <- with(matches, rand.Bet.OTH + rdiscrete(n = nrow(matches), values = 0:3, probs=c(p0, p1, p2, p3)))

  #For the random bets, we have 'placed bets' on the 05/06 season, we will ignore it for the plots
  for(i in n:N){
    if (matches$FTR[i] == "H"){matches$rand.Ret.OTH[i] <- (matches$BbAvH[i]-1) * matches$rand.Bet.OTH[i]}
    else{matches$rand.Ret.OTH[i] <- -matches$rand.Bet.OTH[i]}}

  matches$rand.CumR.OTH <- with(matches, 0)
  matches$rand.CumR.OTH[n] <- matches$rand.Ret.OTH[n]
  for (i in n:N){matches$rand.CumR.OTH[i] <- matches$rand.CumR.OTH[i-1] + matches$rand.Ret.OTH[i]}

  lines(matches$BetIndex, matches$rand.CumR.OTH, col = alpha("blue", alpha0), type = 'l')
}

lines(matches$BetIndex, matches$C.OTHReturns, col = alpha("red"), type = 'l')
lines(matches$BetIndex, matches$C.OTHRet.alt, col = alpha("green"), type = 'l')
legend(0, -2500, c('Our Method', 'Alternate', 'Random'), lty=c(1,1,1), col=c('red', 'green', 'blue'))

#1X2 Away
plot(matches$BetIndex, matches$C.OTAReturns, col = alpha("red"), type = 'l', ylim = c(-3000, 10), xlab = 'Index', ylab = 'Returns', main = '1X2 Away Win')
for (i in 1:nRuns){
  p1 <- nrow(matches[matches$OTAwayBet == 1,]) / nrow(matches[(n+1):N,])
  p2 <- nrow(matches[matches$OTAwayBet == 2,]) / nrow(matches[(n+1):N,])
  p3 <- nrow(matches[matches$OTAwayBet == 3,]) / nrow(matches[(n+1):N,])
  p0 <- 1 - (p1 + p2 + p3)

  #Reset the random bet, 1X2 Away
  matches$rand.Bet.OTA <- with(matches, 0)
  matches$rand.Ret.OTA <- with(matches, 0)
```

```
422   matches$rand.Bet.OTA <- with(matches, rand.Bet.OTA + rdiscrete(n = nrow(matches), values
          = 0:3, probs=c(p0, p1, p2, p3)))
423
424   for (i in n:N){
425     if (matches$FTR[i] == "A"){matches$rand.Ret.OTA[i] <- (matches$BbAvA[i]-1) *
            matches$rand.Bet.OTA[i]}
426     else{matches$rand.Ret.OTA[i] <- -matches$rand.Bet.OTA[i]}}
427
428   matches$rand.CumR.OTA <- with(matches, 0)
429   matches$rand.CumR.OTA[n] <- matches$rand.Ret.OTA[n]
430   for (i in n:N){matches$rand.CumR.OTA[i] <- matches$rand.CumR.OTA[i-1] +
          matches$rand.Ret.OTA[i]}
431
432   lines(matches$BetIndex, matches$rand.CumR.OTA, col=alpha("blue", alpha0), type='l')
433 }
434
435 lines(matches$BetIndex, matches$C.OTAReturns, col = alpha("red"), type = 'l')
436 lines(matches$BetIndex, matches$C.OTARet.alt, col = alpha("green"), type = 'l')
437 legend(0, -2500, c('Our Method', 'Alternate', 'Random'), lty=c(1,1,1), col=c('red',
        'green', 'blue'))
438
439 #AH Home
440 plot(matches$BetIndex, matches$C.AHHReturns, col = alpha("red"), type = 'l', ylim =
        c(-3000, 10), xlab = 'Index', ylab = 'Returns', main = 'AH Home Win')
441 for (i in 1:nRuns){
442   p1 <- nrow(matches[matches$AHHomeBet == 1,]) / nrow(matches[(n+1):N,])
443   p2 <- nrow(matches[matches$AHHomeBet == 2,]) / nrow(matches[(n+1):N,])
444   p3 <- nrow(matches[matches$AHHomeBet == 3,]) / nrow(matches[(n+1):N,])
445   p0 <- 1 - (p1 + p2 + p3)
446
447   #Reset the random bet, AH Home
448   matches$rand.Bet.AHH <- with(matches, 0)
449   matches$rand.Ret.AHH <- with(matches, 0)
450   matches$rand.Bet.AHH <- with(matches, rand.Bet.AHH + rdiscrete(n = nrow(matches), values
          = 0:3, probs=c(p0, p1, p2, p3)))
451   for (i in n:N){
452     if (matches$ah.res[i] == "hm"){matches$rand.Ret.AHH[i] <- (matches$BbAvAHH[i]-1) *
            matches$rand.Bet.AHH[i]}
453     else if (matches$ah.res[i] == "hfhm"){matches$rand.Ret.AHH[i] <- (matches$BbAvAHH[i]-1)
            * 0.5 * matches$rand.Bet.AHH[i] - (0.5 * matches$rand.Bet.AHH[i])}
454     else{matches$rand.Ret.AHH[i] <- -matches$rand.Bet.AHH[i]}}
455
456   matches$rand.CumR.AHH <- with(matches, 0)
457   matches$rand.CumR.AHH[n] <- matches$rand.Ret.AHH[n]
458   for (i in n:N){matches$rand.CumR.AHH[i] <- matches$rand.CumR.AHH[i-1] +
          matches$rand.Ret.AHH[i]}
459
460   lines(matches$BetIndex, matches$rand.CumR.AHH, col=alpha("blue", alpha0), type='l')
461 }
462 lines(matches$BetIndex, matches$C.AHHReturns, col = alpha("red"), type = 'l')
463 lines(matches$BetIndex, matches$C.AHHRet.alt, col = alpha("green"), type = 'l')
464 legend(0, -2500, c('Our Method', 'Alternate', 'Random'), lty=c(1,1,1), col=c('red',
```

```
    'green', 'blue'))

#AH Away
plot(matches$BetIndex, matches$C.AHAReturns, col = alpha("red"), type = 'l', ylim =
    c(-3000, 10), xlab = 'Index', ylab = 'Returns', main = 'AH Away Win')
for (i in 1:nRuns){
  p1 <- nrow(matches[matches$AHAwayBet == 1,]) / nrow(matches[(n+1):N,])
  p2 <- nrow(matches[matches$AHAwayBet == 2,]) / nrow(matches[(n+1):N,])
  p3 <- nrow(matches[matches$AHAwayBet == 3,]) / nrow(matches[(n+1):N,])
  p0 <- 1 - (p1 + p2 + p3)

  #Reset the random bet, AH Away
  matches$rand.Bet.AHA <- with(matches, 0)
  matches$rand.Ret.AHA <- with(matches, 0)
  matches$rand.Bet.AHA <- with(matches, rand.Bet.AHA + rdiscrete(n = nrow(matches), values
      = 0:3, probs=c(p0, p1, p2, p3)))
  for (i in n:N){
    if (matches$ah.res[i] == "aw"){matches$rand.Ret.AHA[i] <- (matches$BbAvAHA[i]-1) *
        matches$rand.Bet.AHA[i]}
    else if (matches$ah.res[i] == "hfaw"){matches$rand.Ret.AHA[i] <- (matches$BbAvAHA[i]-1)
        * 0.5 * matches$rand.Bet.AHA[i] - (0.5 * matches$rand.Bet.AHA[i])}
    else{matches$rand.Ret.AHA[i] <- -matches$rand.Bet.AHA[i]}}

  matches$rand.CumR.AHA <- with(matches, 0)
  matches$rand.CumR.AHA[n] <- matches$rand.Ret.AHA[n]
  for (i in n:N){matches$rand.CumR.AHA[i] <- matches$rand.CumR.AHA[i-1] +
      matches$rand.Ret.AHA[i]}

  lines(matches$BetIndex, matches$rand.CumR.AHA, col=alpha("blue", alpha0), type='l')
}
lines(matches$BetIndex, matches$C.AHAReturns, col = alpha("red"), type = 'l')
lines(matches$BetIndex, matches$C.AHARet.alt, col = alpha("green"), type = 'l')
legend(0, -2500, c('Our Method', 'Alternate', 'Random'), lty=c(1,1,1), col=c('red',
    'green', 'blue'))


par(mfrow = c(1, 1)) #Reset graphical parameters
#- End -
```

# Appendix E

# Random Bet Strategy Runs

Tables for the *random bet strategy* in Section 4.4, ran ten times, with an average taken found.

For each run, the following variables are used as the column headings, in order:

- rand**xx** the **xx**$^{\text{th}}$ run of the RBS.
- bets The number of bets placed.
- stake The number of units staked.
- wins The total winnings.
- prop The proportion return: defined as the winnings divided by the stake.
- diff The difference in the proportion return, between the RBS and Table 4.2.
- acc (pc) The accuracy (% of bets won), displayed as a percentage.
- diff The difference in the accuracy, between the RBS and Table 4.2.

| rand01 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19651 | 39544 | -2270.173 | -0.057 | -0.028 | 44.247 | -19.116 |
| 1x2 A | 20048 | 40286 | -3244.41 | -0.081 | -0.049 | 30.163 | -17.310 |
| AH H | 11680 | 20872 | -4086.35 | -0.196 | -0.147 | 40.899 | -4.705 |
| AH A | 19673 | 28306 | -5116.955 | -0.181 | -0.105 | 40.151 | -3.473 |
| Overall | 71052 | 129008 | -14717.888 | -0.114 | -0.071 | 38.589 | -11.920 |

| rand02 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19630 | 39331 | -2103.088 | -0.053 | -0.024 | 44.101 | -19.262 |
| 1x2 A | 19855 | 39833 | -4295.35 | -0.108 | -0.076 | 29.519 | -17.953 |
| AH H | 11700 | 21003 | -3810.35 | -0.181 | -0.133 | 41.171 | -4.433 |
| AH A | 19585 | 28306 | -5075.6 | -0.179 | -0.104 | 40.521 | -3.104 |
| Overall | 70770 | 128473 | -15284.388 | -0.119 | -0.076 | 38.535 | -11.974 |

| rand03 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19801 | 39780 | -1859.475 | -0.047 | -0.017 | 44.447 | -18.915 |
| 1x2 A | 19991 | 40444 | -4436.47 | -0.110 | -0.078 | 29.058 | -18.414 |
| AH H | 11677 | 20873 | -3357.865 | -0.161 | -0.112 | 42.468 | -3.136 |
| AH A | 19936 | 28704 | -5018.245 | -0.175 | -0.099 | 40.449 | -3.175 |
| Overall | 71405 | 129801 | -14672.055 | -0.113 | -0.070 | 38.699 | -11.810 |

| rand04 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19735 | 39616 | -2226.513 | -0.056 | -0.026 | 44.322 | -19.040 |
| 1x2 A | 20170 | 40774 | -3708.15 | -0.091 | -0.059 | 30.045 | -17.428 |
| AH H | 11714 | 20791 | -4091.21 | -0.197 | -0.148 | 41.071 | -4.534 |
| AH A | 19813 | 28435 | -4827.835 | -0.170 | -0.094 | 40.514 | -3.111 |
| Overall | 71432 | 129616 | -14853.708 | -0.115 | -0.071 | 38.701 | -11.808 |

| rand05 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19628 | 39383 | -2475.863 | -0.063 | -0.033 | 44.207 | -19.155 |
| 1x2 A | 19933 | 39919 | -2365.53 | -0.059 | -0.028 | 30.296 | -17.176 |
| AH H | 11759 | 20959 | -3829.69 | -0.183 | -0.134 | 41.577 | -4.027 |
| AH A | 19440 | 28045 | -5111.305 | -0.182 | -0.107 | 40.108 | -3.517 |
| Overall | 70760 | 128306 | -13782.388 | -0.107 | -0.064 | 38.725 | -11.784 |

| rand06 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19634 | 39340 | -2821.68 | -0.072 | -0.042 | 43.613 | -19.750 |
| 1x2 A | 20292 | 40962 | -2302.37 | -0.056 | -0.025 | 30.337 | -17.135 |
| AH H | 11775 | 21071 | -3662.955 | -0.174 | -0.125 | 41.308 | -4.296 |
| AH A | 19700 | 28428 | -5113.81 | -0.180 | -0.104 | 40.416 | -3.208 |
| Overall | 71401 | 129801 | -13900.815 | -0.107 | -0.064 | 38.578 | -11.931 |

| rand07 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19564 | 39302 | -2566.27 | -0.065 | -0.035 | 44.137 | -19.226 |
| 1x2 A | 19880 | 40033 | -2760.17 | -0.069 | -0.037 | 30.226 | -17.246 |
| AH H | 11695 | 20983 | -3883.53 | -0.185 | -0.136 | 41.214 | -4.390 |
| AH A | 19600 | 28142 | -4827.445 | -0.172 | -0.096 | 40.378 | -3.247 |
| Overall | 70739 | 128460 | -14037.415 | -0.109 | -0.066 | 38.703 | -11.806 |

| rand08 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19676 | 39357 | -2324.514 | -0.059 | -0.029 | 44.415 | -18.948 |
| 1x2 A | 20003 | 40293 | -2731.89 | -0.068 | -0.036 | 30.380 | -17.092 |
| AH H | 11775 | 21040 | -3557.045 | -0.169 | -0.120 | 41.622 | -3.982 |
| AH A | 19803 | 28373 | -5049.685 | -0.178 | -0.103 | 40.181 | -3.444 |
| Overall | 71257 | 129063 | -13663.134 | -0.106 | -0.062 | 38.837 | -11.672 |

| rand09 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19632 | 39463 | -2396.331 | -0.061 | -0.031 | 43.979 | -19.384 |
| 1x2 A | 20001 | 40105 | -3021.75 | -0.075 | -0.044 | 29.924 | -17.549 |
| AH H | 11638 | 20948 | -3469.505 | -0.166 | -0.117 | 41.622 | -3.982 |
| AH A | 19716 | 28224 | -4980.005 | -0.176 | -0.101 | 40.277 | -3.348 |
| Overall | 70987 | 128740 | -13867.591 | -0.108 | -0.064 | 38.604 | -11.905 |

| rand10 | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19570 | 39202 | -1845.215 | -0.047 | -0.017 | 44.844 | -18.519 |
| 1x2 A | 19953 | 40197 | -2956.07 | -0.074 | -0.042 | 30.437 | -17.036 |
| AH H | 11820 | 21329 | -3601.305 | -0.169 | -0.120 | 41.878 | -3.726 |
| AH A | 19740 | 28342 | -5068.93 | -0.179 | -0.103 | 40.198 | -3.427 |
| Overall | 71083 | 129070 | -13471.52 | -0.104 | -0.061 | 39.016 | -11.492 |

Taking the average of each cell:

| avgs | bets | stake | wins | prop | diff | acc (pc) | diff |
|---|---|---|---|---|---|---|---|
| 1x2 H | 19652.1 | 39431.8 | -2288.9122 | -0.058 | -0.028 | 44.231 | -19.131 |
| 1x2 A | 20012.6 | 40284.6 | -3182.216 | -0.079 | -0.047 | 30.038 | -17.434 |
| AH H | 11723.3 | 20986.9 | -3734.9805 | -0.178 | -0.129 | 41.483 | -4.121 |
| AH A | 19700.6 | 28330.5 | -5018.9815 | -0.177 | -0.102 | 40.319 | -3.305 |
| Overall | 71088.6 | 129033.8 | -14225.0902 | -0.110 | -0.067 | 38.699 | -11.810 |

# Appendix F

# Project Diary

## Meeting One — 25/09/20

Maha, Alun and I discussed the start of my project: how it will be assessed, what I need to do each week, ETC. We spoke about where the data can be found (`football-data.co.uk`), how to import it, how to use it, and how it is formatted. We went over the basics of probabilities and their relationships with odds and the bookmaker's commissions (typically 5%). Before next week, I will read the Kaunitz, Zhong, and Kreiner (2017) and look at replicating the simple steps (means, standard deviations) with my data.

## Meeting Two — 02/10/20

We reviewed the Kaunitz, Zhong, and Kreiner (2017) paper, and looked at my R code used to complete the first steps. Before next week, I need to find out how `football-data.co.uk` source their data; do exploratory work on R (such as histograms and other basic plots) for either the 1X2 market across multiple seasons and leagues, or look at goals data for one league.

## Meeting Three — 09/10/20

After Meeting Two, I made plots of observed vs. bookmaker probabilities using fixed points (this ended up with some very small bins): we discussed how it would be better to make sure the bin sizes are equal, instead. Maha spoke about the need for ensuring my code is well-commented, and to start properly writing up what graphs show (it will make my final write up easier). Finally, we realised `football-data.co.uk` renamed their `BbAvH` column to `AvgH`: I had come across problems with using data from multiple seasons.

## Meeting Four — 16/10/20

We looked into problems I faced with previous code, caused by small sample sizes. The point was made that `tapply` can be used to find the mean of a bin, rather than taking the midpoint, creating far more accurate plots and linear models. We discussed methods for comparing accuracy over time (statistics such as $R^2$, mean squared error, ETC. or tests such as Kendall's Tau or Pearson's Rank),

and spoke about the affect of *competitive balance*, and how we can quantify it to use in part of our analysis.

## Meeting Five — 23/10/20

With Alun, we spoke about what plots I have created showed, and how I can use them in my project as a point for discussion. We discussed why RMSE is a better measure of accuracy than mean square error, median square error or absolute square error. Alun also showed me some extra papers, such as Owen (2009), with methods and statistics ($P_1$ and $P_2$) that can be applied to the project. We discussed the future of the project, and how I will continue to progress. Finally, we spoke about the need to record everything I do, and to include more significant figures in my values for $R^2$.

## Meeting Six — 06/11/20

Up until now, I had imported the `football-data.co.uk` datasets individually: Maha showed me a much better method using `for` loops and the `paste0` R command, which cut down a 1,000+ line document to around 100. We spoke about a number of papers I found about competitive balance, and the coefficients they used to quantify it. We spoke about using a binomial regression model, using $1/O_i$ (inverse of odds) as a predictor, and adding the season and leagues as additional ones, to see what affect they have. Finally, we spoke about what information would be included in my exploratory data analysis section.

## Meeting Seven — 13/11/20

We spoke about using an ordinal logistic model, rather than binomial, and the pros and cons of each (OLM is simpler to code and run, but we may have multicollinearity between Home Win and Away Win probabilities). Alun demonstrated how the `ordinal` package worked, with the `clm` and `clmm` commands. We discussed whether or not the country would be treated as a fixed factor, and finally, we spoke about looking into how often the bookmakers 'favourite' won, and different ways we could define this.

## Meeting Eight — 20/11/20

Maha gave me information about the literature review, and how I could prepare for this: for each paper I use, write down a summary of the paper, including:

- what the paper is about,
- how the researchers found the results (their method),
- the results they showed,
- and similarities and differences with my paper.

Maha recommended I create a table with these headings to fill in each time I use a new paper. We also discussed my ethical approval (the first check is on December 7$^{\text{th}}$ 2020).

## Meeting Nine — 27/11/20

Due to coursework deadlines, I hadn't done much new work on my project, so we went over my progress so far, and had a discussion about what I've shown so far. We concluded I'm making good headway into the project, and how I can improve my results. We discussed other packages and techniques I can use, such as the `erer`, `marginal` and `mass` packages.

## Meeting Ten — 16/12/20

Throughout the last week, I'd been working through a number of papers and summarising them (as per Maha's suggestion, Meeting Eight). We discussed what I need to show by the January check. I'd also made a number of plots using the `ggplot2` package, which Alun and I went over, making a number of comments about them. We looked at changing the parameters for the density plot (default is Gaussian kernel estimate, and $n = 512$). We discussed these plots at length and how the matches with a high probability of a Draw are likely due to match fixing or when a Draw suits both teams. At length, we discussed the difference in density plots for each league, and why some leagues appeared to have a unimodal distribution, and others having a trimodal distribution. Alun suggested using Odds Portal to check any unusual odds. We also discussed a tile plot I created, and discussed that grouping 5+ goals was key: small bin sizes are not helpful. I also gave a brief outline of possible sections for my first section, which we modified, and what progress I'm expected to make over Christmas. Finally, we discussed the Under/Over 2.5 Goals market, and the Asian Handicap market, and how I could look into the accuracy of these over the Christmas break, and throughout Semester Two.

## Meeting Eleven — 18/01/21

### Code and Content

Over the Christmas break, I'd made a section of code to find $R^2$, RMSE, $P_1$, and $P_2$ for each league that was 205 lines long: we discussed how a `for` loop would be much better, like how we did to read in the data. For the tile plot, we discussed possible ways of including bin size: we both agreed inserting a table below would be clearest. Maha confirmed a question I had about including code in-text, rather than referring to it in the appendices.

### Write-Up

In my submitted document, I couldn't get references to work: Maha recommended the `natbib` LaTeX package, and we spoke about how APA and Harvard are the best referencing systems. Maha also made it clear figures should be in the middle of the text, where they are needed in context, so the reader doesn't have to continually refer back and forth. A small conclusion at the end of each chapter is required, as well as a final concluding chapter with our findings and discussion. Figure labels need to be below the figure; table labels above. Maha also sent an annotated copy of my write-up.

## Meeting Twelve — 28/01/21

As well as making changes based on our previous meeting, I'd found a good paper discussing different leagues styles of play: as a result, I conducted principal components analysis with three new variables:

`predAcc` (predictive accuracy), `imbalance` (a measure of the competitive imbalance in a league), and `attack` (a measure based on the shots per goal: more shots implies a more attacking league). The Kruskal test is a good addition, as well as potentially running PCA on the seasons, rather than leagues. Finally, after Maha's comments on my draft about normalising the probabilities, I found Shin's Method of Normalisation (Štrumbelj, 2014), which I will apply and compare over the next week.

## Meeting Thirteen — 16/02/21

In the two weeks, I finalised my *elite* data analysis, and moved onto the English & Scottish leagues, including the Under/Over 2.5 Goals and Asian Handicap markets. The aforementioned Shin's Method was excluded, due to incorrect findings of consensus probabilities. Due to Maha being off ill, I caught Alun up with my progress, to which he made a few comments: I should look into the leverage of my models, especially with the Under/Over 2.5 Goals market; the $\chi^2$ test for independence between goals is good, but due to some expected values being low, I would need to group goals together: it also doesn't take into account the order of goals, so perhaps a different test (Spearman or Pearson) is better. Alun also introduced me to the Skellam distribution, based on the Poisson distribution, as a way to model goals, and recommended I read about bivariate Poisson distributions.

## Meeting Fourteen — 02/03/21

I made large amounts of progress in the two weeks, and started a new section, creating a betting model based on bookmaker's odds. Alun suggested making a few changes, such as changing the range of my bets for different markets and changing the units I bet on. We discussed ignoring the lower performing leagues and markets (such as German Bundesliga and Level 2 in the English & Scottish leagues, and the U/O market).

## Meeting Fifteen — 16/03/21

Alun and I went through all of my outputs (figures, tables, ETC.), and discussed what each one showed. He recommended I talk about the amount of coding knowledge I have acquired throughout the dissertation, and challenges I have faced throughout the course of the project, and how I have overcame them. We also discussed the presentation, and how I can both prepare for it, and what to include in it.

## Meeting Sixteen — 30/03/21

Alun and I discussed the final stages of the project. We went through my presentation slides, suggesting possible areas of improvement, and what needs to be included in my script. We went through what will be completed by the second check deadline, and what I will submit then, as well as a discussion into whether or not it is necessary to include the correlation tests in the dissertation (Kendall's Tau, Spearman Rank, and Goodman-Kruskal-Gamma tests): adding something with no value can make the dissertation harder to read. Finally, we spoke about the conclusion chapter, namely the inclusion of common themes and contradictory findings in it.

## Meeting Seventeen — 13/04/21

After a discussion about the presentation (when it would happen, what to include, ETC.), Alun and I went through my April draft, looking at mistakes and things to chance, such as going into more depth at points, altering the Structure paragraph and discussing why I included certain steps. Finally, we concluded with a brief look at the progress I have made in the Model chapter, and what progress I need to make before submission.

## Meeting Eighteen — 20/04/21

Alun reviewed my draft of the Literature Review and Abstract, and going over my mistakes, such as discussing the importance of not referring to 'we' in the Abstract. We then spoke about the presentation for a final time, ensuring the technology works and what to expect from the two markers.

# Appendix G

# Word Count

Full word count: **20,307 words**

**Breakdown:**

| *Chapter* | *Prose* | *Figures* | *Tables* | *Algorithms* | **Total** |
|---|---|---|---|---|---|
| Abstract | 270 | | | | 270 |
| 1: Introduction | 3377 | | 246 | | 3623 |
| 2: *Elite* Leagues | 6095 | 110 | 679 | | 6884 |
| 3: English & Scottish Pyramids | 4113 | 301 | 887 | | 5301 |
| 4: Proposed Method | 1882 | 142 | 258 | 462 | 2744 |
| 5: Conclusion | 1485 | | | | 1485 |
| **Total** | 17222 | 553 | 2070 | 462 | **20307** |

# Appendix H

# Ethical Approval Certificate



## Certificate of Ethical Approval

| | |
|---|---|
| Applicant: | Joseph Pym |
| Project Title: | Assessing the Accuracy of Betting Odds in Football |

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Medium Risk

| | |
|---|---|
| Date of approval: | 14 Dec 2020 |
| Project Reference Number: | P114620 |