CAN MATHS BEAT THE BOOKIES?

BRADLEY MAUNDER

Abstract. This paper focuses on predicting UK football results with a view towards creating a betting strategy. Two of the most basic models in the literature are used to calculate attack and defence parameters for teams in the 2010/2011 Premier League. These parameters are then used to predict probabilities of football results in the 2011/2012 Premier League season to form a betting strategy. This betting strategy yields a positive return when using the maximum odds available for each game at a discrepancy level of 1.1.

Contents

1. Introduction

Is it possible to predict the outcome of football matches given previous results? Is it possible to form a betting strategy that is superior enough to beat the bookmakers? This paper looks to present the current literature on this topic, reproduce some of the most influential models, create a hypothetical betting strategy and answer these questions.

1.1. **History.** Betting on football outcomes may have unofficially been around for just as long as football itself. With the introduction of online betting websites, mobile applications and the half time television advertisements, bookmaker's profits are booming [23] and betting on sports outcomes is more popular than ever. Some bookmakers also offer the in-play betting market. This allows the punter to bet on certain outcomes of the match whilst it is still being played. A selection of bookmakers also allow the punter to cash out their earnings before the match has finished. Technology has revolutionised the betting market and it is thus very popular.

1.2. **Background Knowledge.** The main approach of this paper is to use mathematics of a statistical nature. Some of the methods used feature maximum likelihood estimation and the difference of two means, among various other statistical techniques. Many of these techniques are revised in Section 2. Most of the mathematical content used is of degree level or less, but for a full understanding or revision of the statistical methods that are not included in this paper we recommend *Introduction to mathematical statistics (7th edn)* [17]. However any general statistics book of degree level would suffice. Some references have been used, where needed, to suggest further reading for a better understanding/interest.

The statistical software 'R' is also used and when commands are presented there will be annotations explaining the method. R can be freely downloaded for different operating systems online at *http://http://cran.r-project.org/bin/* [24].

1.3. **Applications.** Many models have been produced that could benefit both the average punter and also the bookmakers. The mathematical models discussed in this paper predict football match results, or even match scores and can therefore form the basis of a betting strategy. R codes can be copied from the appendices to allow the reader to apply the betting strategies discussed in this paper to their

own data. Alternatively visit *https://github.com/maunderb/Betting-Strategy* [29] to access the codes. Also, as the models produce probabilities of a home win, draw, or away win, match odds for bookmakers could be suggested.

1.4. **Literature Review.** There has been much research and development of mathematical models that predict football results in recent years however the most cited seems to be the model produced by Maher in 1982 [1]. Maher used a model that assumed home and away team's scores were from a Poisson distribution and were independent of each other. This model allowed teams to have different attacking and defencive strengths when playing at home or away. He then proposed a second model that assumed a team had only one set of attacking and defencive strengths and that these were simply weighted down when playing away. From comparing these two models he concluded that a model with single attack and defence parameters for home and away was the better option. He also found a good fit when applying this model to the original data. However, Maher did suggest that a possible improvement of his model could be to use a bivariate Poisson model thus incorporating dependence between home and away scores. Subsequently Dixon and Coles [2] extended Mahers model to include this dependence between scoring rates. This model was then further improved by introducing a weighting that allowed recent results to be more influential than results from, say, five years ago. This allowed fluctuating performances to affect their predictions. Dixon and Coles then used their model as a basis for a betting strategy by predicting match result probabilities which when used against bookmaker's odds produced a positive return. Karlis and Ntzoufras [7] also discussed the use of bivariate Poisson models and Poisson difference models in sport. They concluded that such models proved very useful and accurate when used with the right data.

There have also been other approaches to building mathematical models that predict football results. For example, some models have focused on the timings of the goals scored. Dixon and Robinson set up a two-dimensional birth process that varied with time to predict the timings of goals scored by home and away teams during a match [4], which again was based upon Maher's [1]. Meanwhile, Crowder, Dixon, Ledford and Robinson used the idea of a bivariate stochastic process to refine Dixon and Cole's model which allowed the attacking and defensive strengths of teams to evolve through time [6]. Alternatively Goddard focused on comparing the models that used data of goals scored to models that used data of match results to determine how well each predicted match outcomes[10]. He found no statistically significant difference between the ability of the two models to predict match outcomes; but rather that a model using both results and goals of previous games seemed to be the best. Also, Owen used Dynamic generalised linear models (DGLMs) to allow for a time variation of parameters [12].

Other papers have focused on the efficiency of the UK association football betting market. As early as 1989 Pope and Peel [3], using a model based on Maher's

[1], concluded that there did not appear to be a profitable betting strategy yielding a positive return. Also, Cain, Law and Peel [5], who focused on the long-shot bias - the phenomenon that bookmakers overvalue long-shots and under value favourites, concluded that a long-shot bias does in fact exist and also that both Poisson and negative binomial models allow good estimation of results but do not offer significant betting opportunities. In 2004 Dixon and Pope compared different bookmaker's odds and predictions to a model similar to Dixon and Coles [2], and decided that the market was inefficient [9]. In particular a reverse long-shot bias was found. Meanwhile Graham and Stott [11] developed an Ordered Probit model to compare predictive results and bookmakers odds.

In 2009 Vlastakis, Dotsis and Markellos [15] focused on the efficiency of the European football betting market. Among their findings were plenty of arbitrage opportunities allowing the market to be exploited by particular betting strategies. They found that a long-shot bias existed and they also produced Poisson count models and multinomial logit regression models that better predicted football results and scores than betting odds. A more recent paper used a Bayesian network model to predict football outcomes which also produced a profit when used as a betting strategy (Constantinou, Fenton and Neil) [13]. They took a different route and ranked teams at each game according to a predetermined rank table. They found that subjective information improved the accuracy of the model. In contrast to results of Dixon and Pope they found a long-shot bias against bettors, as opposed to a reverse long-shot bias. Meanwhile Koopman and Lit produced a model that used a bivariate Poisson distribution which allowed for coefficients to change randomly over time [14]. Again, this produced a positive return when used to bet against the bookmakers.

1.5. **Aim.** There are various ways to bet on a football match; the result, the correct score, whether there will be over/under 2.5 goals, the first goal scorer, the number of yellow/red cards issued and many more. The most popular of these however is the result (a home win, a draw or an away win). It is for that reason that we focus on the result of a football match and present a model that tries to predict just that. We also focus on the over or under 2.5 goals betting market, the reason for this being that as far as we are aware this market is not discussed or used in the current literature.

This paper first introduces the basic model used by Maher and then applies it to data from the 2010/2011 English Premier League to predict results of the 2011/2012 Premier League season in Section 3. We then present a second model used by Maher in Section 4 and compare these two models. Using the results of the second model we then find the probabilities of results and the probabilities of seeing over/under 2.5 goals in individual games sampled from the 2011/2012 season. Using these probabilities we set up a betting strategy to reveal whether a positive return can be made when compared to bookmakers' odds from 2011/2012, this forms Section 5. Our findings are then discussed in Section 6.

1.6. **Data.** There is a vast amount of data readily available on previous football results as all statistics of a match are recorded. Some examples include the score, the number of corners, the number of shots, and the possession of the ball held by each team. Primarily we are interested in the main statistic: the goals scored. All data used in this paper is extracted from the football-data website [22]. Historical data of over a dozen football seasons can be downloaded in excel format. We will focus on the Premier League division (the top football league in England) and only focus on the previous years worth of results (corresponding to 2010/11) due to the simplicity of the model - see Section 4.5 for more details.

In the Premier League there are 20 teams that all play each other twice, once at home and once away. Therefore there are ${}^{20}P_2 = 380$ matches in a season. During any one season a team may also participate in league and cup games but these are ignored for simplicity. As we are interested in predicting the outcomes of football matches from the 2011/2012 season we only wanted to focus on the results of the 20 teams that were in the Premier League during this particular season. Although, of these 20 teams not all would have been in the Premier League the previous year. This is due to three teams being relegated to the lower league (Championship) and three teams being promoted from the Championship to the Premier League. However producing a model that can predict match results for the newly promoted teams proves difficult (see Section 4.5) and it is for this reason that we only focus on the 17 teams that remained in the Premier League from the 2010/2011 to 2011/2012 season.

Data for the maximum odds are also available in the spreadsheets extracted from the football-data website. Historical odds of a home win, draw, or an away win are available and also the odds of over/under 2.5 goals being scored are available. This information will be useful when forming our betting strategy in Section 5.

The total goals scored and the number of goals conceded at Home for the Premier League season 2010/2011 for each of the 20 teams can be seen in Figure 1.

2. Preliminaries

In this section we introduce some of the basic statistical definitions that are needed to understand most of the methods used in this paper.

Definition 1. Let X be a random variable and $\lambda \in \mathbb{R}_+$. The Poisson distribution is defined [18, p. 1] by

$$\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots.$$

We write $X \sim \text{Poisson}(\lambda)$ (which reads: X is distributed by the Poisson distribution with parameter λ).

Definition 2. Let $X_1, X_2, ..., X_n$ be discrete random variables. If observed values are $x_1, x_2, \ldots, x_n$ then the likelihood function L for a discrete random variable is

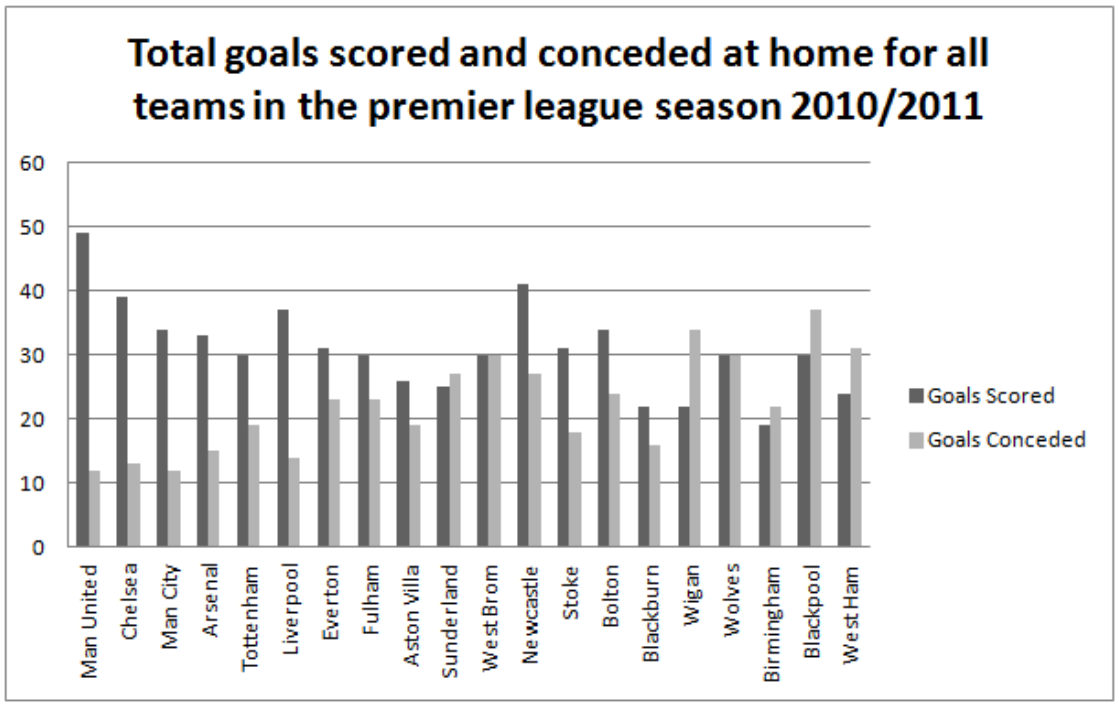

FIGURE 1. Goals scored and goals conceded at home for the season 2010/2011. Teams are in order of final league position from left to right.

[19, p. 59]

$$L(x_1, x_2, ..., x_n) = \Pr\left[\cap_{j=1}^{n} (X_j = x_j \mid \theta_1, \theta_2, \ldots, \theta_m)\right].$$

Where $\theta_1, \theta_2, \ldots, \theta_m$ are parameters of the random variables $X_1, X_2, \ldots, X_m$. The *log-likelihood* is simply $l = \log(L(x_1, x_2, \ldots, x_n))$.

Definition 3. Maximising the likelihood can usually be achieved by solving the equations

$$\frac{\partial L\left(x_1, x_2, \ldots, x_n \mid \theta_1, \theta_2, \ldots, \theta_m\right)}{\partial \theta_i} = 0, \quad \text{for } i = 1, 2, \ldots, m$$

and solving for θ_i, will give you your $\hat{\theta}_i$.

The values $\hat{\theta_1}, ..., \hat{\theta_m}$ are called *maximum likelihood estimators*.

The log-likelihood will always be negative or zero, thus maximising the likelihood is usually equivalent to maximising the log-likelihood.

The maximum log-likelihood is $l\left(x_1, x_2, \ldots, x_n \mid \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta_m}\right)$, i.e. the log-likelihood with its maximum likelihood estimators substituted in to it[19, p. 59].

3. Model 1

The following model is the first model introduced by Maher [1]. Let X_{ij} and Y_{ij} be random variables. We make the following assumptions:

(1) Each game has a score of (x_{ij}, y_{ij}) where x_{ij} is the number of goals scored by team i against team j and y_{ij} is the number of goals scored by team j against team i.
(2) Team i is playing at home and team j is playing away.
(3)
$$X_{ij} \sim \text{Poisson}\left(\alpha_i \beta_j\right)$$
and
$$Y_{ij} \sim \text{Poisson}\left(\gamma_i \delta_j\right),$$
where α_i can be viewed as the home team's attacking strength, β_j the away team's defencive strength, γ_i the home team's defencive strength and δ_j the away team's attacking strength.
(4) X_{ij} and Y_{ij} are independent, see Section 4.5.

We now present the estimation of the parameters here, it can be applied to any football league with n teams. For example, as there are 20 teams in the Premier League there are 80 parameters to be estimated for the Premier League.

We now derive the log-likelihood function which will lead us to estimate our parameters. By Definition 1 it is easy to see that

$$Pr\left(x; \alpha, \beta\right) = \frac{\left(\alpha_i \beta_j\right)^{x_{ij}} e^{-\alpha_i \beta_j}}{\left(x_{ij}\right)!}.$$

Finding the likelihood and applying the log function, we can see that by Definition 2, the log-likelihood l is

$$l\left(x;\alpha, \beta\right) = \sum_{i}^{n} \sum_{j \neq i}^{n} \left(x_{ij} \log\left(\alpha_i \beta_j\right) - \log\left(x_{ij}!\right) - \alpha_i \beta_j\right).$$

Now, we take the partial derivative with respect to α_i giving

$$\frac{\partial}{\partial \alpha_i} l\left(x_{ij}; \alpha_i, \beta_j\right) = \sum_{j \neq i}^{n} \left(-\beta_j + \frac{x_{ij}}{\alpha_i}\right).$$

If we set this equal to zero and solve for α_i then

$$\sum_{j \neq i}^{n} \left(-\beta_j + \frac{x_{ij}}{\alpha_i}\right) = 0 \iff \alpha_i = \frac{\sum_{j \neq i}^{n} x_{ij}}{\sum_{j \neq i}^{n} \beta_j}.$$

Similarly if we take the partial derivative with respect to β_j of the above log-likelihood, set this equal to zero and solve for β_j we can see that

$$\sum_{i\neq j}^{n}\left(-\beta_j+\frac{x_{ij}}{\alpha_i}\right)=0 \iff \beta_j=\frac{\sum_{i\neq j}^{n}x_{ij}}{\sum_{i\neq j}^{n}\alpha_i}.$$

Our alpha and beta parameters to be estimated therefore satisfy

$$\hat{\alpha}_i=\frac{\sum_{j\neq i}^{n}x_{ij}}{\sum_{j\neq i}^{n}\hat{\beta}_j}, \hat{\beta}_j=\frac{\sum_{i\neq j}^{n}x_{ij}}{\sum_{i\neq j}^{n}\hat{\alpha}_i}. \tag{1}$$

The same method is used to find an equation for the gamma and delta parameters. Simply replace α_i by γ_i, β_j by δ_j and x_{ij} by y_{ij} in the calculations above to find:

$$\hat{\gamma}_i=\frac{\sum_{j\neq i}^{n}y_{ij}}{\sum_{j\neq i}^{n}\hat{\delta}_j}, \hat{\delta}_j=\frac{\sum_{i\neq j}^{n}y_{ij}}{\sum_{i\neq j}^{n}\hat{\gamma}_i}. \tag{2}$$

See Appendix A or visit *https://github.com/maunderb/Betting-Strategy* [29] for how these parameters can be estimated in R. The codes are not given by Maher [1] and we therefore provide our own codes that can be used to reproduce the results by the reader. A list of the estimated parameters for Model 1 can be seen in Table 1 .

4. Model 2

4.1. **Maher's improvement.** Whether all of these parameters are needed could be questioned. In Maher [1], a hierarchy of models was introduced to test this question. Maher suggested that attack and defence strengths are not needed for both home and away matches, but rather that attack and defence strengths differ between home and away matches by a factor, say k. That is, in a match where the score is (x_{ij}, y_{ij}), team j's attack and team i's defence strengths are weighted down by a factor k; this is the same for all teams $i \in [1, n]$. This suggestion is supported by likelihood ratio tests in Maher's paper that test the need for all the parameters in Model 1. We now present Model 2 from Maher's paper and, as the derivation for the likelihood is omitted in Maher's paper, we show it here in more detail.

As in Model 1 (see Section 3) it is assumed that

$$X_{ij}\sim \text{Poisson}\left(\alpha_i\beta_j\right),$$

however in a different manner if we let $\delta_j = k\alpha_j$ and $\gamma_i = k\beta_i$ then

$$Y_{ij}\sim \text{Poisson}\left(k^2\alpha_j\beta_i\right),$$

where α_i is the strength of team i's attack, β_j is the strength of team j's defence and k is a constant parameter to be estimated.

TABLE 1. Estimated Parameters for Model 1

Model 1	**Parameter Estimates**			
Team i	$\hat{\alpha}_i$	$\hat{\beta}_i$	$\hat{\gamma}_i$	$\hat{\delta}_i$
Arsenal	1.3239531	1.190527	0.7397533	1.9134497
Aston Villa	1.0641035	1.6821674	0.9003662	1.0879551
Birmingham	0.7717203	1.4955613	1.0331329	0.8960281
Blackburn	0.9046202	1.7962824	0.7614799	1.1787641
Blackpool	1.2305417	1.7363752	1.7704552	1.2919006
Bolton	1.3736041	1.3634807	1.1272774	0.9002471
Chelsea	1.5440976	0.8584114	0.6272109	1.4638018
Everton	1.2309328	0.9317289	1.0852997	0.9981789
Fulham	1.1871586	0.8454589	1.0827437	0.948149
Liverpool	1.4900582	1.2846375	0.6630311	1.0753341
Man City	1.3480257	0.8938101	0.573472	1.2653105
Man United	1.958163	1.0924288	0.577509	1.4115852
Newcastle	1.6517414	1.2935936	1.2596058	0.755204
Stoke	1.2477643	1.2714458	0.8391243	0.7395478
Sunderland	1.0040409	1.2164986	1.2746202	1.0077004
Tottenham	1.2012574	1.1420502	0.9067578	1.2367035
West Brom	1.2305417	1.7363752	1.4374955	1.3208501
West Ham	0.980331	1.6343554	1.4606031	0.966371
Wigan	0.880383	1.1267574	1.5986083	0.9219809
Wolves	1.2199116	1.523936	1.4032421	0.8114188

The maximum likelihood estimates in Model 1 were

$$\hat{\alpha}_i = \frac{\sum_{j\neq i}^{n} x_{ij}}{\sum_{j\neq i}^{n} \hat{\beta}_j}, \tag{3}$$

$$\hat{\beta}_j = \frac{\sum_{j\neq i}^{n} x_{ij}}{\sum_{j\neq i}^{n} \hat{\alpha}_i}, \tag{4}$$

and these are exactly the same for X_{ij}. However, as there are alphas and betas in Y_{ij}, the maximum likelihood estimates need to satisfy two more equations (5 and 6). Replacing $\hat{\gamma}_i$ by $\hat{k}\hat{\beta}_i$ and $\hat{\delta}_j$ by $\hat{k}\hat{\alpha}_j$ in equation 2 we get

$$\hat{k}\hat{\alpha}_j = \frac{\sum_{i\neq j}^{n} y_{ij}}{\sum_{i\neq j}^{n} \hat{k}\hat{\beta}_i} \iff \hat{k^2}\hat{\alpha}_j = \frac{\sum_{i\neq j}^{n} y_{ij}}{\sum_{i\neq j}^{n} \hat{\beta}_i} \iff \hat{k^2}\hat{\alpha}_i = \frac{\sum_{j\neq i}^{n} y_{ji}}{\sum_{j\neq i}^{n} \hat{\beta}_j} \tag{5}$$

and

$$\hat{k}\hat{\beta}_i = \frac{\sum_{j\neq i}^{n} y_{ij}}{\sum_{j\neq i}^{n} \hat{k}\hat{\alpha}_j} \iff \hat{k^2}\hat{\beta}_i = \frac{\sum_{j\neq i}^{n} y_{ij}}{\sum_{j\neq i}^{n} \hat{\alpha}_j} \iff \hat{k^2}\hat{\beta}_j = \frac{\sum_{i\neq j^n}^{n} y_{ji}}{\sum_{i\neq j^n} \hat{\alpha}_i}, \tag{6}$$

where we have multiplied by $\hat{k}$ on both sides and then swapped the indices i and j.

If we compute the sums (equation 3 + equation 5) and (equation 4 + equation 6), with some rearranging, we get the following maximum likelihood estimates for α_i and β_j

$$\hat{\alpha_i} = \frac{\sum_{j\neq i}^{n}(x_{ij}+y_{ji})}{\left(1+\hat{k^2}\right)\sum_{j\neq i}^{n}\hat{\beta_j}}$$

and

$$\hat{\beta_j} = \frac{\sum_{i\neq j}^{n}(x_{ij}+y_{ji})}{\left(1+\hat{k^2}\right)\sum_{i\neq j}^{n}\hat{\alpha_i}}.$$

We also find that

$$\hat{k}^2 = \frac{\sum_i \sum_{j\neq i} y_{ij}}{\sum_i \sum_{j\neq i} x_{ij}}.$$

These estimates are consistent with the maximum likelihood estimates stated in Maher [1].

See Appendix B or visit *https://github.com/maunderb/Betting-Strategy* for how these parameters can be estimated in R. Again we have supplied the R codes in a format so that the reader can estimate their own estimated parameters using different data.

A list of the estimated parameters for Model 2 can be seen in Table 2.

4.2. **The Home Advantage.** It seems logical that a team playing at their home ground would have a slight advantage over their opponents. After all, they are playing on a pitch that they are very familiar with, they have their home crowd to cheer them on and they haven't had to endure a long trip to get there. The theory that there is a so called 'Home Advantage' is supported by our results and also much of the literature.

The mean of all the 380 home teams' scores in the premier league 2010/2011 season is 1.62 and the mean of all the 380 away teams' scores only 1.17; this alone suggests there is some advantage. However to construct a more formal statistical argument we present a test of the difference of two means. The method used can be seen in *Introduction to mathematical statistics (7th edn)* [17, p. 219-220].

Let μ_1 be the population mean for the home teams' scores and μ_2 be the population mean for the away teams' scores, estimated by $\bar{x}_1$ and $\bar{x}_2$. Let n_1 and n_2 be the number of observations for the home teams' and away teams' scores respectively. We also assume that s_1 and s_2 are the sample standard deviations (estimates of the population standard deviations σ_1 and σ_2) of the home teams' and away teams' scores respectively. Let us assume that the two samples are independent, that $n_1 = n_2 = 380 > 30$ and that $\sigma_1 = \sigma_2$. We would like to test the hypotheses:

$$H_0 : \mu_1 = \mu_2$$

TABLE 2. Estimated Parameters for Model 2

Model 2	**Parameter Estimates**	
Team i	$\hat{\alpha}_i$	$\hat{\beta}_i$
Arsenal	1.7546744	1.0770011
Aston Villa	1.1871903	1.4441055
Birmingham	0.9136487	1.4042205
Blackburn	1.1375918	1.4412379
Blackpool	1.3873443	1.9246093
Bolton	1.282579	1.3759414
Chelsea	1.6647072	0.8234945
Everton	1.2442343	1.1039626
Fulham	1.1929969	1.0527289
Liverpool	1.438486	1.0879271
Man City	1.4471559	0.8162321
Man United	1.8899438	0.9318965
Newcastle	1.3829301	1.4061943
Stoke	1.1253099	1.1719563
Sunderland	1.109497	1.3664177
Tottenham	1.3433816	1.1330117
West Brom	1.4026162	1.7529706
West Ham	1.0749215	1.7056638
Wigan	0.9908227	1.4813889
Wolves	1.145555	1.6127464
$\hat{k}^2 =$	0.7228525	

$$H_1 : \mu_1 \neq \mu_2.$$

From R we have $\bar{x}_1 = 1.623684$, $\bar{x}_2 = 1.173684$, $s_1 = 1.240464$ and $s_2 = 1.114236$.

The pooled sample standard deviation is

$$s_p \coloneqq \sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}} = 1.179041.$$

Now $\sigma_{\bar{x}_1 - \bar{x}_2}$ is estimated by

$$s_{\bar{x}_1 - \bar{x}_2} \coloneqq s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 0.0855366.$$

Assuming H_0 is true, the test statistic t is

$$t \coloneqq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}} = 5.260906.$$

The critical value is calculated in R:

```
qt(0.995,n1+n2-2)

[1] 2.582331
```

As $t = 5.26 > 2.58$ we reject H_0 and can conclude that there is evidence to suggest a significant difference between the two means at the 1% significance level (in fact this is significant at the 0.00001% level).

As mentioned in the previous subsection, Maher [1] introduced a home ground advantage and weighted the away teams' attacking and home teams' defensive strengths down by a constant amount, say k, an amount that does not vary from team to team. Of course this home ground advantage can also be captured by Model 1. It seems that almost all of the literature reviewed in Section 1.4 uses a home advantage factor in their models. Clarke and Norman [16], for example, wrote a paper that focused entirely on the home advantage in football matches. They showed that home advantage effects played a significant part in predicting the final result of matches, and that there was some variation in these factors from team to team. However these differences were not statistically significant. Graham and Stott [11] also confirmed that the home advantage effect was not statistically different between different teams. Hirotsu and Wright [8] further studied different factors affecting football results, one of which was the home advantage. They found that it was an important factor. This suggests that the home advantage effect plays some part in mathematical models that predict football results and thus supports the use of Model 1 and Model 2.

4.3. **Model Comparison.** In Section 4.1 we introduced a second model that required less parameters to be estimated. In this section we show that this model is a good model by means of Akaike's Information Criterion (AIC).

Definition 4. The AIC of a particular model is defined [20, p. 57] as:

$$AIC = -2 \times (\text{max log likelihood of model}) + 2 \times (\text{no. of free parameters model}).$$

When comparing two or more models the model with the smallest AIC is preferred.

We now have maximum likelihood estimates for the parameters of Model 1 and Model 2 and can therefore calculate their respective maximum log-likelihoods. To compare these two models we do not have to look at the log-likelihoods for X_{ij} because although the estimates of α_i and β_j will be different between the models, the maximum log-likelihoods are approximately the same:

$$l\left(x;\hat{\alpha},\hat{\beta}\right)=-640.9352$$

and

$$l\left(x;\hat{k}^2,\hat{\alpha},\hat{\beta}\right)=-641.4764.$$

We therefore focus on the log-likelihoods for Y_{ij} because these *are* different. The log-likelihood function for Y_{ij} in Model 1 is

$$l\left(y;\gamma,\delta\right)=\sum_{i}^{n}\sum_{j\neq i}^{n}\left(y_{ij}\log\left(\gamma_i\delta_j\right)-\log\left(y_{ij}!\right)-\gamma_i\delta_j\right)$$

and the log-likelihood function for Y_{ij} in Model 2 is

$$l\left(y;k^2,\alpha,\beta\right)=\sum_{i}^{n}\sum_{j\neq i}^{n}\left(y_{ij}\log\left(k^2\alpha_j\beta_i\right)-\log\left(y_{ij}!\right)-k^2\alpha_j\beta_i\right).$$

If we replace the parameters in these likelihood functions by their maximum likelihood estimates given in Sections 3 and 4.1 we arrive at the following maximum likelihood functions:

$$l\left(y;\hat{\gamma},\hat{\delta}\right)=-500.6688$$

and

$$l\left(y;\hat{k}^2,\hat{\alpha},\hat{\beta}\right)=-514.2364.$$

Let us denote the number of free parameters needed for each Y_{ij}, as k_1 and k_2 for model 1 and 2 respectively. It can be seen by Section 3 that $k_1 = 40$ and, as the only extra parameter needed to be estimated for Y_{ij} in Model 2 is k^2, $k_2 = 1$. This is because if we have estimated all the α's and β's from X_{ij} then all we need to estimate all the means of Y_{ij}'s is the k^2. We can then compute $AIC_{Model1} = 1081.338$ and $AIC_{Model2} = 1030.473$. As the AIC for model 2 is smaller than that of model 1, we decide to use model 2 (as Maher does).

4.4. **Goodness of fit.** We now have our preferred model, but is this model a good fit to the data?

Following Maher [1], in this subsection we conduct a Chi-squared goodness of fit test. Using our parameter estimates for the premier league teams in 2010/2011 from Section 4.1 we can estimate the probability of gaining the scores (x_{ij}, y_{ij}) for all $i, j \in [1, 20]$ and $x, y \in [0, 4+]$ for Model 2. We can then find the expected score frequencies. To do this we do the following:

(1) Calculate the estimated parameters of X_{ij} and Y_{ij} for all $i, j \in [1, 20]$ for Model 2 from the maximum likelihood estimates in Table 2. Remember that $X_{ij} \sim \text{Poisson}\left(\alpha_i\beta_j\right)$ and $Y_{ij} \sim \text{Poisson}\left(k^2\alpha_j\beta_i\right)$ in Model 2.
(2) Using R, for every score in a match (x_{ij}, y_{ij}) for all $i, j \in [1, 20]$ find the probabilities of gaining a score of $(0, 0), (1, 0), (0, 1), \ldots, (4+, 4+)$ (for details on how to calculate these probabilities please see Definition 6).

(3) Add up the total of these probabilities of gaining $0, 1, 2, 3, 4+$ goals scored by the home team and the total of the probabilities of gaining $0, 1, 2, 3, 4+$ goals scored by the away for each of the 380 matches. These probabilities can also be seen as the expected values in one game.
(4) Sum these 380 different expected values to find the expected score frequencies over one full season (380 games).

Let us now give an example to illustrate this method.

Example 1. Let us focus on one of the 380 games: Everton vs Tottenham. We would like to find the expected frequencies for $0, 1, 2, 3, 4+$ goals scored at home and away. We follow the numbering above:

(1) Everton is team number 8 and Tottenham is team number 16. So from Table 2 we can see that $X_{8,16} \sim$ Poisson (1.409732) and $Y_{8,16} \sim$ Poisson (1.072021).
(2) From R we calculate the following probability matrix:

$$P = \begin{pmatrix} \mathbf{0.08360} & 0.08962 & 0.04804 & 0.01717 & 0.00579 \\ 0.11785 & 0.12634 & 0.06772 & 0.02420 & 0.00817 \\ 0.08307 & 0.08905 & 0.04773 & 0.01706 & \mathbf{0.00576} \\ 0.03903 & 0.04185 & 0.02243 & 0.00802 & 0.00271 \\ 0.01877 & 0.02012 & 0.01078 & 0.00385 & 0.00130 \end{pmatrix}.$$

The rows correspond to the goals $0, 1, 2, 3, 4+$ scored for the home team and the columns correspond to the goals $0, 1, 2, 3, 4+$ scored for the away team. For example, the probability of gaining a score of $(0, 0)$ is 0.08360 and the probability of gaining a score of $(2, 4)$ is 0.00576. These values have been made bold.

These probabilities can also be seen as the expected score frequencies from 1 game. If these teams played each other 100 times you would expect $100 \times 0.12634 \approx 13$ to have a score of $(1, 1)$ and $100 \times 0.02243 \approx 2$ to have a score of $(3, 2)$.
(3) We now sum the rows to find the total number of expected home teams' scores and sum the columns to find the total number of expected away teams' scores for this particular match. These can be seen in Table 3.
(4) We would have 380 of these expected frequencies and they would be summed to give the expected score frequencies of the whole season.

We estimate the expected frequency of all the 380 scores and categorise them as represented in Table 4.4. The actual (Observed) scores are categorised in the same manner.

Before conducting any tests, if we observe the Plots 2 and 3, we can make some judgment on how well Model 2 fits the data. Plots 2 and 3 are histograms for the observed Home and Away goals scored, with Model 2's expected frequency line superimposed on top. In terms of home scores (Plot 2) Model 2 tends to underestimate the amount of $1, 2$ goals scored and over-estimate the goals greater than

TABLE 3. Expected score frequency of a Match between Everton and Tottenham

	Game 1	
Goals	**Home**	**Away**
0	0.244	0.342
1	0.344	0.367
2	0.243	0.197
3	0.114	0.070
4+	0.055	0.024
Total	1	1

TABLE 4. Observed and expected scores for home and away matches for Model 2

Model 2	**Home**		**Away**		**Home**	**Away**
Goals	**Observed (O)**	**Expected (E)**	**O**	**E**	$\frac{(O-E)^2}{E}$	
0	65	77.64204	126	117.96997	2.058436066	0.546591491
1	131	115.11754	126	131.06593	2.191260651	0.195807154
2	108	92.58086	78	79.15024	2.568024086	0.016715705
3	50	53.76355	39	34.5556	0.263455605	0.571620558
4+	26	40.89601	11	17.25826	5.425739917	2.269395537
Total	380	380	380	380	12.50691633	3.600130446

or equal to 3. This is consistent with the results that Maher [1] found. However Model 2 is still a fairly reasonable fit to the data. When looking at away scores (Plot 3) Model 2 appears to fit the data very well.

We can now formalise these arguments in the form of a Chi-Squared goodness of fit test. We propose the following hypotheses:

H_0 : Model 2 is a good fit to the Home (Away) teams' scores data,

H_1 : Model 2 is not a good fit to the Home (Away) teams' scores data.

Note that there are actually two sets of hypotheses here.

To test these hypotheses we need to calculate the test statistic $\chi^2 = \sum_i \frac{(O-E)^2}{E}$ which is approximately distributed by the χ^2 distribution with 3 degrees of freedom. To see why this is the case please see Maher [1, p. 114]. We would reject H_0 if the test statistic χ^2 is greater than $\chi^2_3(0.05)$ (the 0.95 quantile of the χ^2 distribution with 3 degrees of freedom).

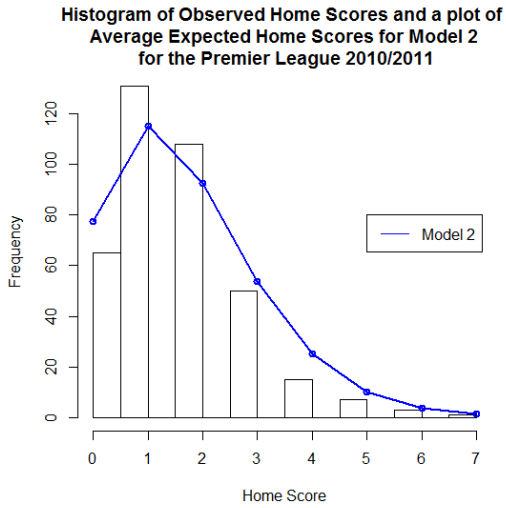

FIGURE 2. A histogram of observed home scores and a plot of average expected home scores of Model 2.

In our test $\chi^2_3(0.05) = 7.814728$. Our test statistics can be extracted from the bottom right of Table 4.4. For Model 2, the home χ^2 value is 12.5069, which does lead to a rejection of H_0 and the χ^2 value for the away scores is 3.6001, which does not reject H_0. This suggests that the home teams' score estimations *are not* a very good fit to the data whereas the away teams' score estimations *are* a good fit to the data.

From these tests we can conclude that Model 2 fits the data well for away teams' scores, but not so well for home teams' scores. However we can still regard the Poisson model as a good model however it's results must be interpreted with caution. Our results are therefore slightly different to Maher [1], in that Maher finds Model 2 a good fit to the data on most occasions.

4.5. **Limitations.** In the following sections we only proceed to predict results for 17 of the 20 teams in the Premier League season 2011/2012 thus limiting the number of games we can bet on from 380 to 272. This is due to the fact that three of the teams in the season 2011/2012 were in the Championship the previous year (the league below) and gained promotion to the Premier League. Parameters could have been estimated for these teams and therefore their results predicted, however as they would have been playing against teams of a lower quality and thus probably have won more games in 2010/2011, the parameters would not be representative of their future performance in the higher (i.e. tougher) league. This limitation is not ideal and some models do eliminate this problem by incorporating previous results of cup games, where teams from lower leagues would have played teams in the higher leagues. Dixon and Coles' model incorporated these cup games and therefore overcame this limitation. However, in only being able to predict 108

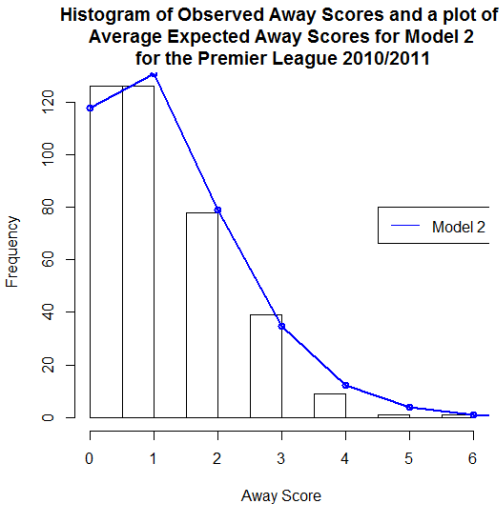

FIGURE 3. A histogram of observed away scores and a plot of average expected away scores of Model 2.

games less per season we still believe that our model is useful when used as a betting strategy and this can be seen in Section 5.

Another limitation of our Model 2 is that we only use results from one season. This is again due to the movement of teams via promotion and relegation. Although, we suggest that results from more than 2 or 3 years ago can be seen as unrepresentative of a teams future performance and therefore may not provide a good enough prediction anyway. This was supported by Dixon and Coles [2] from that they introduced a weighting function that allowed for fluctuating performances and allowed more weight on recent games.

A further limitation is based on our assumption that home teams' and away teams' scores are independent. If you are familiar with football, this idea seems incomprehensible. For example if one team scores it may urge the other to 'up their game' and push for an equaliser. Dixon and Coles [2] allowed for a dependence parameter in their improved model which eliminated this unrealistic assumption. Again, although this assumption that we make is unrealistic, we still produce a good model and subsequently a great betting strategy so we keep the independence for simplicity.

5. BETTING STRATEGY

5.1. **Applications of the parameters.** Now that we have our model and parameter estimates, what can we do with it all?

Suppose we have two teams playing in an upcoming match: Manchester United vs Bolton Wanderers. We would like to know the probabilities of either team winning or both drawing, the most likely score and whether there will be over/under a

total of 2.5 goals in the game. We now list some definitions which are immediately followed by an example relating to the above match. This will help us with our betting strategy in the Subsections 5.7 and 5.8.

Remember that in Model 2 $X_{ij} \sim \text{Poisson}(\alpha_i \beta_j)$ and $Y_{ij} \sim \text{Poisson}(k^2 \alpha_j \beta_i)$.

The following notation is similar to the notation used in Dixon and Coles [2].

Definition 5. The probability of a home win between teams i and j is

$$P_{ij}^H := \sum_{x>y}^{\infty} \Pr(X_{ij} = x) \times \Pr(Y_{ij} = y),$$

the probability of an away win is

$$P_{ij}^A := \sum_{x<y}^{\infty} \Pr(X_{ij} = x) \times \Pr(Y_{ij} = y)$$

and the probability of a draw is

$$P_{ij}^D := \sum_{x=y}^{\infty} \Pr(X_{ij} = x) \times \Pr(Y_{ij} = y).$$

Definition 6. The probability of seeing a particular score of (x_p, y_p) between teams i and j is

$$P_{ij}^{(x_p,y_p)} := \Pr(X_{ij} = x_p) \times \Pr(Y_{ij} = y_p).$$

Definition 7. The probability of the total number of goals in a match between teams i and j exceeding 2.5 is

$$P_{ij}^{Over} := \sum_{(x+y)>2.5}^{\infty} \Pr(X_{ij} = x) \times \Pr(Y_{ij} = y)$$

and the probability of the total number of goals being below 2.5 is

$$P_{ij}^{Under} := \sum_{(x+y)<2.5}^{\infty} \Pr(X_{ij} = x) \times \Pr(Y_{ij} = y).$$

Example 2. Manchester United corresponds to team number 12 and Bolton corresponds to team number 6. We can see from Table 1 that $\hat{\alpha_{12}} = 1.8899438$, $\hat{\beta_{12}} = 0.9318965$, $\hat{\alpha_6} = 1.2825790$, $\hat{\beta_6} = 1.3759414$ and $k^2 = 0.7228525$. From this we can see that $X_{12,6} \sim \text{Poisson}(2.600452)$ and $Y_{12,6} \sim \text{Poisson}(0.8639756)$.

Therefore, according to model 2, the Definitions 5, 6 and 7, and using 'R' we can calculate the following probabilities:

$$P_{12,6}^H := \sum_{x>y}^{\infty} \Pr(X_{12,6} = x) \times \Pr(Y_{12,6} = y) = 0.746,$$

$$P_{12,6}^{A} \coloneqq \sum_{x<y}^{\infty} \Pr\left(X_{12,6}=x\right) \times \Pr\left(Y_{12,6}=y\right) = 0.152,$$

$$P_{12,6}^{D} \coloneqq \sum_{x=y}^{\infty} \Pr\left(X_{12,6}=x\right) \times \Pr\left(Y_{12,6}=y\right) = 0.102,$$

$$P_{12,6}^{(3,1)} \coloneqq \Pr\left(X_{12,6}=3\right) \times \Pr\left(Y_{12,6}=1\right) = 0.106,$$

$$P_{12,6}^{Over} \coloneqq \sum_{(x+y)>2.5}^{\infty} \Pr\left(X_{12,6}=x\right) \times \Pr\left(Y_{12,6}=y\right) = 0.673$$

and

$$P_{12,6}^{Under} \coloneqq \sum_{(x+y)<2.5}^{\infty} \Pr\left(X_{12,6}=x\right) \times \Pr\left(Y_{12,6}=y\right) = 0.327.$$

We can see that our model make Manchester United overwhelming favourites with the probability of a win at 74.6%, the probability of seeing a score of $3:1$ at 10.6% and the probability of there being over a total of 2.5 goals in the game at 67.3%.

See Appendix C or visit *https://github.com/maunderb/Betting-Strategy* [29] for details on how to obtain these probabilities for yourself using R. There is also this particular example's R output in the Appendix C.

5.2. **Bookmakers odds.** In the UK, bookmaker's odds are often stated as a/b, for example, 9/2 (often read as "9 to 2" and referred to as UK or fractional odds format) meaning that if you were to place £2 on this particular bet you would receive a return of £9 if your bet wins, a total of £11 including your stake. However elsewhere in the world, for example most other countries in Europe, this bet could be seen as 4.5 (EU or decimal odds format) [25]. This simply means that if you place £1 on a particular outcome you would receive £4.5 in return if you win, a total of £5.5 including your stake. If you are given UK odds of a/b this can be converted to EU odds by simply calculating $\frac{a}{b}$. Both odds formats are used in this section.

To calculate the bookmaker's probabilities we need the following definition.

Definition 8. Given the UK odds of a home win on a particular match, between teams i and j of a/b, the bookmaker's probability is defined [2] as

$$B_{ij}^{H} \coloneqq \frac{b}{a+b}.$$

Similarly we will denote the bookmaker's probabilities of an away win, draw, correct score and over/under 2.5 goals as $B_{ij}^{A}, B_{ij}^{D}, B_{ij}^{(x_p,y_p)}, B_{ij}^{Over}$ and B_{ij}^{Under} respectively.

5.3. **Bookmakers' Margin.** Given the arbitrary odds of a home win, draw, or away win as 10/3, 9/2, 1/2 respectively we can calculate the probability of each outcome by Definition 8. The probabilities are therefore 0.231, 0.182, 0.667, and as you can see $0.231 + 0.182 + 0.667 = 1.08$. Now, these probabilities should add up to one, however the extra 0.08 is known as the bookmakers' margin [9]. This is to give the bookmaker the edge and ensure they make profits. Seems unfair, however if they didn't there could be a lot of arbitrage opportunities , see Section 5.4.

5.4. **Arbitrage Betting.** One betting strategy a punter might take is only betting on arbitrage opportunities. Arbitrage betting is a set of bets that a punter places where he/she is guaranteed to win irrespective of the outcome [3], sounds good right? Arbitrage opportunities like these really do arise in the betting world and some websites even broadcast them. I visited www.oddsportal.com [26]and a particular match in the Iceland league cup between Fjolnir and Olafsvik had odds of Fjolnir winning at 7/2 from Bookmaker A, odds of the match ending in a draw at 33/10 from Bookmaker B and odds of Olafsvik winning at 23/20 from Bookmaker C. If we add up the probabilities of the outcomes from these odds, (see Section 5.2 on how) we can see that there is an arbitrage opportunity:

$$\frac{2}{2+7} + \frac{10}{10+33} + \frac{20}{20+23} = 0.9199.$$

This doesn't have a bookmaker's margin, but rather a 'punter's margin', if you like. In that if we were to place a £2.23 bet on Fjolnir winning with bookmaker A, a £2.33 bet on a draw with bookmaker B and a £4.66 bet on Olafsvik winning with bookmaker C we would be guaranteed a return of £10 irrespective of the outcome, having only placed £9.22 in total, cashing in a guaranteed profit of 78p! Obviously if these bets were multiplied by 100 we would be guaranteed a profit of £78 , however you would have had to bet a huge total of £922. But if you are guaranteed to win, why wouldn't you? Well, quite often these bets, often referred to as 'sure bets' [27], can be a mistake by the bookmaker and they can, under their terms and conditions [28], void one of the bets. This would then leave you having placed two other bets with two other bookmakers and you no longer having all outcomes covered. Thus you would have no arbitrage opportunity and you risk losing the money you had spent on the other two bets. Therefore this is not so risk free after all.

5.5. **Value betting.** Another betting strategy one might take is value betting. Value betting is betting on a particular outcome that you feel the bookmaker has undervalued [27]. For example, say Team A are playing against Team B and a bookmaker has the odds for a home win, draw, and away win as 13/8, 11/5, 7/4 respectively. This would mean that the bookmaker thinks the probabilities of each outcome are 0.381, 0.313, 0.364. Of course these probabilities can not be viewed as real probabilities because they do not add up to one, this is the bookmaker's

margin, see Section 5.3. Ignoring the bookmaker's margin, if for whatever reason, we believe that Team B have a probability of 0.6 of winning then the bookmaker has under-valued this outcome. A punter may only bet on particular outcomes that have been under-valued by a certain amount, we see more of this in the Sections 5.7 and 5.8.

5.6. **Bookmakers vs Model 2.** We have plotted our probabilities against the bookmakers' probabilities in Figure 4. A diagonal line of $y = x$ has been plotted to highlight by how much our probabilities exceed the bookmakers'; it suggests there are a lot of under-valued odds. All dots above the line $y = x$ correspond to under-valued odds and all dots below the line correspond to over-valued odds, according to our model 2. The values to the left of the graph can be seen as the favourites, as the odds are very small and the values more to the right of the graph can be seen as the long-shots as the odds are larger. It appears that more of the values on the left side of the graph are above the $y = x$ and more of the values on the right side of the graph are below the $y = x$ line. This would suggest that there could be a long-shot bias present in these bookmaker's odds, in that favourites are under-valued and long-shots are over-valued. More tests would of course be needed to verify the validity of this statement. This would however be consistent with Cain, Law and Peel [5], Vlastakis, Dotsis and Markellos [15], and Constantinou, Fenton and Neil [13] who all found that long-shot biases existed. Although, it would be in contrast to Dixon and Pope [9] who found a reverse long-shot bias.

5.7. **Strategy for Result betting.** Betting data for the Premier League in the season 2011/2012 downloaded from football-data.co.uk [22] is used in this section. In particular, the data used are the maximum odds for a home win, draw, and away win for each of the 272 matches, and the maximum odds for over/under 2.5 goals for each of the 272 matches as well as the results of each match.

We adopted a similar approach as Dixon and Coles [2] in that our betting strategy involved betting on a match where our estimated probabilities divided by the bookmaker's probabilities $\frac{P_{ij}}{B_{ij}}$ are greater than some number, say v. That is, we only bet on events that our model predicts the bookmaker has undervalued (Value betting, Section 5.5).

However we only bet on 272 of the 380 games in the 2011/2012 season. See Section 4.5 for more information.

Our strategy is as follows:

(1) Using the parameters from Section 4.1 and the method in Definition 5 we calculate probabilities for all the results in the 272 different matches.
(2) We then calculate the bookmaker's probabilities of each outcome in each match using the method from Definition 5.2.

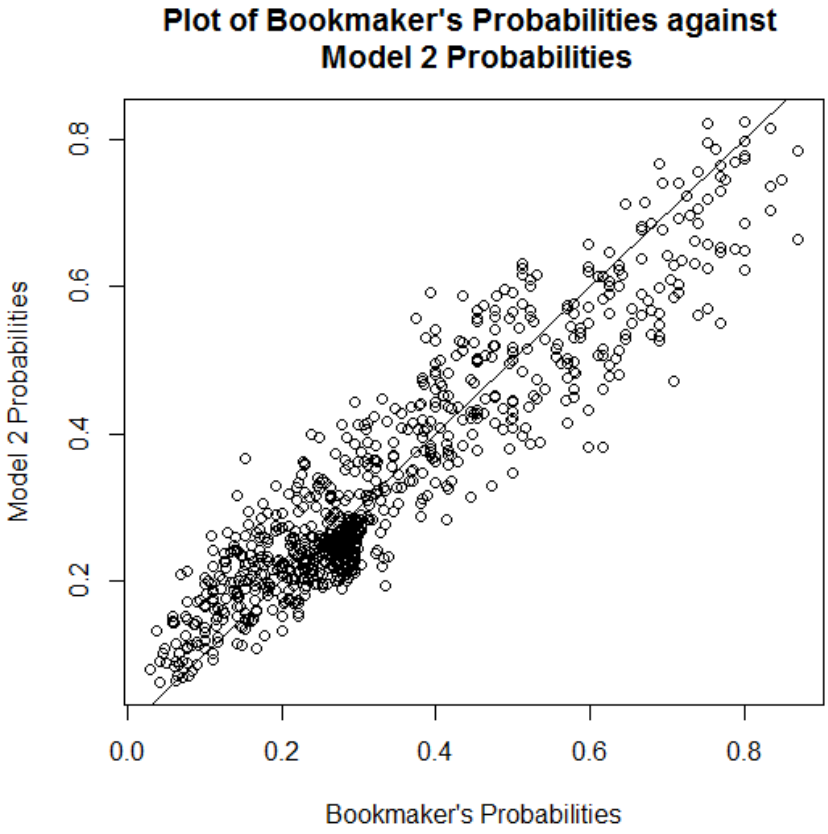

FIGURE 4. Bookmaker's Probabilities of results against Model 2 predicted probabilities

(3) The value $V_{ij} := \frac{P_{ij}}{B_{ij}}$ of each bet is calculated and compared to a predetermined value v. If the value is higher than v, we place a £1 bet on this outcome, otherwise we do nothing.

Example 3. Let $v = 1.1$ and consider a match between Manchester United and Blackburn rovers. We calculate our result probabilities to be $P^H_{12,2} = 0.78437002$, $P^D_{12,2} = 0.13616452$ and $P^A_{12,2} = 0.07946546$. The bookmaker's probabilities are calculated to be $B^H_{12,2} = 0.86956522, B^D_{12,2} = 0.1005025$ and $B^A_{12,2} = 0.03030303$. The values $V_{12,2}$ are therefore $(0.9020255, 1.3548371, 2.6223602)$ for home, draw and away results, and as $2.6223602 > 1.1$ and $1.3548371 > 1.1$ we place a £1 bet on Blackburn rovers to win and both teams to draw. This is because our mathematical model suggests that the events Blackburn to win and both teams to draw have been under-valued by the bookmakers. Blackburn rovers won the game, against the odds, by 3 goals to 2. Now, the maximum odds for this outcome

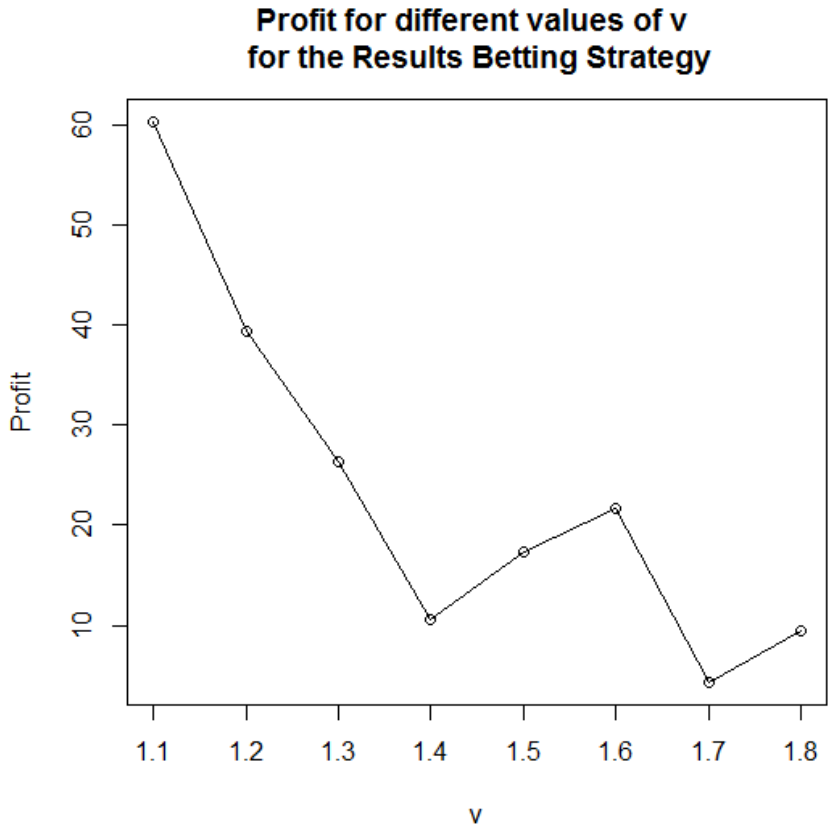

FIGURE 5. Profit for different values of v for Results Strategy

was 33/1, which means that our 2 £1 bets would have won us a return of £34, a profit of £32.

The output for this example in R can be seen in Appendix D. Also, R codes have been provided in Appendix D and at *https://github.com/maunderb/Betting-Strategy* [29] for the reader to make use of this strategy themselves.

We repeated the above strategy for eight different values of v, for all the 272 games and in every single case we would have made a positive return. The highest profit rate being 38.6% and the highest profit amount being £60.23 (a return of £322.23 from £262 worth of bets). Similar to Dixon and Coles [2] who found a choice of $v = 1.2$ to be satisfactory, we found that $v = 1.1$ seemed to give us the greatest return. Obviously as v increases the stricter our betting strategy gets, and fewer bets would be placed. Note we did try values $1.0 < v < 1.1$ however no value gave us a greater return than 1.1. The profit made in each of the different values of v can be seen in Figure 5. As you can see from the graph, the profit generally decreases as v increases.

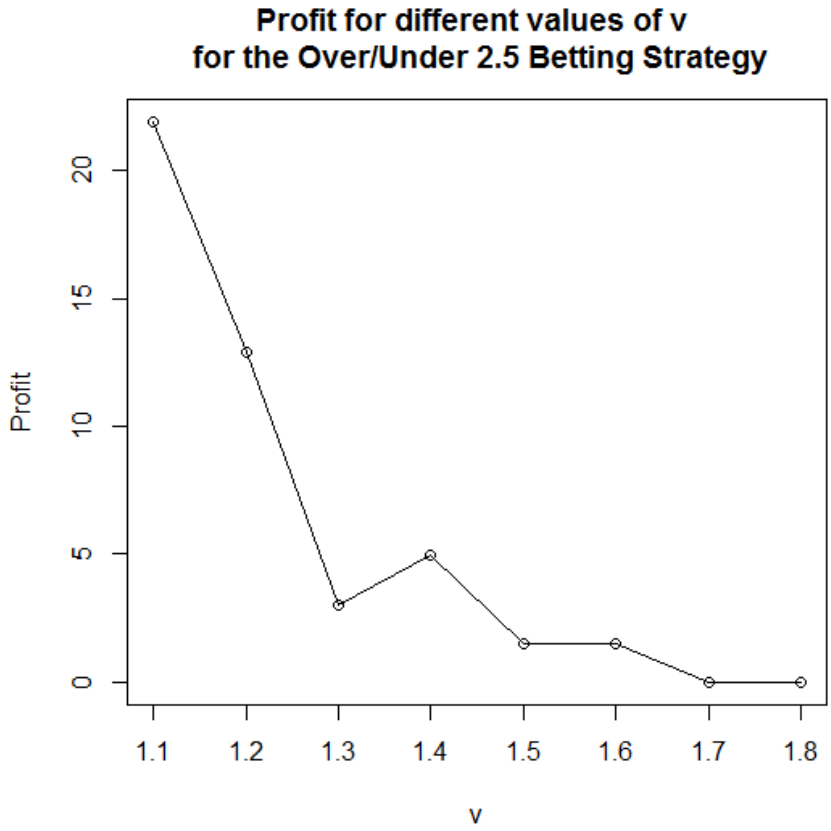

FIGURE 6. Profit for different values of v for Over/Under 2.5 goals Strategy

5.8. **Strategy for over/under 2.5 goals betting.** We also used the previous strategy for the betting market of over/under 2.5 goals. As far as we know a betting strategy for the over/under 2.5 goals betting market has not been implemented in the literature.

Example 4. Let $v = 1.1$ and consider a match between Wolves and Manchester City. We calculate our over/under 2.5 goals probabilities to be $P_{20,11}^{Over} = 0.487$ and $P_{20,11}^{Under} = 0.513$. The bookmaker's probabilities are calculated to be $B_{20,11}^{Over} = 0.667$ and $B_{20,11}^{Over} = 0.333$. The values $V_{20,11}$ are therefore $(0.731, 1.539)$ for Over and Under 2.5 goals, and since $1.539 > 1.1$ we place a £1 bet on there being under 2.5 goals. The score finished $2:0$ to Manchester City and hence there are under 2.5 goals. The maximum odds for this outcome was 3/1, which means that our £1 bet earned us a profit of £3.

The output of this example in R can be seen in Appendix E. Also, R codes have been provided in Appendix E and at *https://github.com/maunderb/Betting-Strategy* [29] for the reader to make use of this strategy themselves.

Again, we repeated this strategy for eight different values of v and in every single case we would have made a positive return. The highest profit rate being 75% and the highest profit amount being £21.90 (a return of £146.90 from £125 worth of bets). Again we did try values $1.0 < v < 1.1$ however no value gave us a greater return than 1.1. Note: the highest profit rate of 75% is based on only two single bets being placed for $v = 1.6$ and thus this value is not considered to be a good betting strategy and should be ignored. The profit made in each of the different values of v can be seen in Figure 6.

6. Discussion

We reproduced the results of two of Maher's [1] models with data from the Premier League 2010/2011. These models were compared and, in agreement with Maher, we concluded that Model 2 was the best model to proceed with. We tested how well this model fitted to the data and found that it fitted well enough to produce fairly accurate predictive probabilities. We then used it to predict probabilities for the results and to whether there would be over/under 2.5 goals in 272 games in the Premier League season 2011/2012. These probabilities were used as the basis of a betting strategy, in that our predicted probabilities were compared to the real bookmaker's probabilities from this season. The betting strategy allowed for bets on discrepancies above a value of 1.1, and thus produced a profit for both the results market and the over/under 2.5 betting market. We also observed that there could be a long-shot bias present in the bookmaker's odds, in that favourites are often under-valued and long-shots over-valued. Again, as this is just an observation from a graph, more tests would be needed to verify the validity of this statement. However if there is a long-shot bias this is consistent with much of the literature.

The aim of this paper was to introduce a mathematical model superior enough to predict results to form the basis of a betting strategy. This objective was achieved and it was surprising that such a basic model introduced by Maher [1] over 30 years ago was able to accurately predict results and form a betting strategy with positive return. It suggests that using an improved model like Dixon and Coles' [2] could produce an even better set of probabilities and subsequently a better betting strategy with higher return.

Possible improvements could include introducing a possession parameter into the model, after all a team can only score a goal when it has possession of the ball. If a team keeps possession particularly well, this would therefore give them some sort of advantage over a team that does not keep the ball so well. This would perhaps give a more realistic and accurate prediction of results. However, since this data was not in an easy to use format, unlike the spreadsheets downloaded from football-data.co.uk [22], this may prove time consuming.

Some of the results in this paper can be reproduced by the reader if he/she wishes by viewing the appendices. The reader can copy and paste the appendices

into R or visit *https://github.com/maunderb/Betting-Strategy* [29], following any instructions provided. The Model did produce a positive return in our case, however, we do not take any responsibility for any losses that a reader may experience as a result of this paper. Please gamble responsibly.

Appendix A. Estimation of Parameters for Model 1 in R

```
# Download your predictive data from www.football-data.co.uk and save as
# "Data.csv" (a comma seperated value file).
# Setting up the data:
DATA=read.csv("Data.csv",header=T) # This reads the data into R
HD=DATA[with(DATA,order(DATA[,3],DATA[,4])),] # Ordering the data by
# Home Team
AD=DATA[with(DATA,order(DATA[,4],DATA[,3])),] # Ordering the data by
# Away Team
# Naming some of the columns:
HOME1=HD[,3]
AWAY1=HD[,4]
HOME2=AD[,3]
AWAY2=AD[,4]
HOME.SCORE1=HD[,5]
AWAY.SCORE1=HD[,6]
HOME.SCORE2=AD[,5]
AWAY.SCORE2=AD[,5]
SH=sum(HD[,'FTHG']) # sums the total home scores
SA=sum(HD[,'FTAG']) # sums the total away scores
facH1=factor(HOME1) # This is useful when indexing by teams
facA1=factor(AWAY1)
facH2=factor(HOME2)
facA2=factor(AWAY2)
f1=function(x)return(sum(x[['FTHG']]))
f2=function(x)return(sum(x[['FTAG']]))
bHH1=by(HD,facH1,f1)
bHH2=by(HD,facH1,f2)
bHA1=by(HD,facA1,f1)
bAH2=by(AD,facH2,f2)
bAA2=by(AD,facA2,f2)
bAA1=by(AD,facA2,f1)
# We now estimate the parameters for Model 1:
# Initial estimates can be calculated as:
A=c(bHH1)/sqrt(SH)
B=c(by(HD,facA1,f1))/sqrt(SH)

# We then iteratively calculate the parameters for all the teams:
B1=c(bHA1)/(sum(A)-A)
A1=c(bHH1)/(sum(B1)-B1)
B2=c(bHA1)/(sum(A1)-A1)
A2=c(bHH1)/(sum(B2)-B2)
```

```
B3=c(bHA1)/(sum(A2)-A2)
A3=c(bHH1)/(sum(B3)-B3)
B4=c(bHA1)/(sum(A3)-A3)
A4=c(bHH1)/(sum(B4)-B4)
B5=c(bHA1)/(sum(A4)-A4)
A5=c(bHH1)/(sum(B5)-B5)

# Also, the same again for the GAMMA and DELTA parameters
G=c(bAH2)/sqrt(SA)
D=c(bAA2)/sqrt(SA)
D1=c(bAA2)/(sum(G)-G)
G1=c(bAH2)/(sum(D1)-D1)
D2=c(bAA2)/(sum(G1)-G1)
G2=c(bAH2)/(sum(D2)-D2)
D3=c(bAA2)/(sum(G2)-G2)
G3=c(bAH2)/(sum(D3)-D3)
D4=c(bAA2)/(sum(G3)-G3)
G4=c(bAH2)/(sum(D4)-D4)
D5=c(bAA2)/(sum(G4)-G4)
G5=c(bAH2)/(sum(D5)-D5)
R1=cbind(A5,B5,G5,D5) # This is a table of parameter estimates for Model 1
```

Appendix B. Estimation of Parameters for Model 2 in R

```
# Model 2
k2=SA/SH # This is the estimate for k^2.

# As in model 1, we make initial estimates:
NAL=c(bHH1+bAA2)/((1+k2)*sqrt(SH))
NBE=c(bAA1+bHH2)/((1+k2)*sqrt(SH))
# and now we iteratively estimate the parameters for Model 2
NBE1=c(bAA1+bHH2)/((1+k2)*(sum(NAL)-NAL))
NAL1=c(bHH1+bAA2)/((1+k2)*(sum(NBE)-NBE))
NBE2=c(bAA1+bHH2)/((1+k2)*(sum(NAL1)-NAL1))
NAL2=c(bHH1+bAA2)/((1+k2)*(sum(NBE1)-NBE1))
NBE3=c(bAA1+bHH2)/((1+k2)*(sum(NAL2)-NAL2))
NAL3=c(bHH1+bAA2)/((1+k2)*(sum(NBE2)-NBE2))
NBE4=c(bAA1+bHH2)/((1+k2)*(sum(NAL3)-NAL3))
NAL4=c(bHH1+bAA2)/((1+k2)*(sum(NBE3)-NBE3))
NBE5=c(bAA1+bHH2)/((1+k2)*(sum(NAL4)-NAL4))
NAL5=c(bHH1+bAA2)/((1+k2)*(sum(NBE4)-NBE4))

R2=cbind(NAL5,NBE5) # A table of the estimated parameters for Model 2
```

Appendix C. Applications of the Parameters

```
# The following functions can be used to estimate probabilities given home and
# away score means, that is, for model 2, for a match between team i and team
# j, p = α̂_i β̂_j and q = k̂² α̂_j β̂_i.

# Please note when copying and pasting the following commands the { signs
# will be replaced by the letter 'f' and the } signs will be replaced by the letter
# 'g'. Unfortunately these will need to be amended before using the codes.

# Probability of results given home and away score means
prob=function(p,q){
hs1=dpois(c(0:1000),p)
as1=dpois(c(0:1000),q)
sc=outer(hs1,as1,"*")
HOME=sum(sc[lower.tri(sc)])
DRAW=sum(diag(sc))
AWAY=sum(sc[upper.tri(sc)])
return(cbind(HOME,DRAW,AWAY))}

# Probability of scores given home and away score means
probsc=function(p,q){
hs1=dpois(c(0:10),p)
as1=dpois(c(0:10),q)
sc=outer(hs1,as1,"*")
return(sc)}

# Prob over/under 2.5 goals given home and away score means
prob2.5=function(p,q){
hs1=dpois(c(0:1000),p)
as1=dpois(c(0:1000),q)
sc=outer(hs1,as1,"*")
over=sum(sc)-(sc[1,1]+sc[1,2]+sc[2,1]
+sc[3,1]+sc[1,3]+sc[2,2])
under=1-over
return(cbind(over,under))}

# For the example in Section 5.1 (Example 2) p = 2.600452 and q = 0.8639756:
# > prob(2.600452,0.8639756)
# HOME      DRAW      AWAY
# [1,] 0.7460317 0.1524547 0.1015136
```

```
# > prob2.5(2.600452,0.8639756)
#    over      under
# [1,] 0.6725232 0.3274768
# These results are the probabilities stated in Section 5.1.
```

Appendix D. Betting Strategy for Results

```
# The following functions can be used to decide whether to bet on particular
# outcomes on a single football match or not. The variables p and q are the same
# as in Appendix C, h, d, a need to be EU odds for home win, draw or away win
# respectively and r is the discrepency level, we reccomend setting r = 1.1. Where
# the R output shows a 0 no bet should be placed, where the output displays a 1
# a bet should be placed on this outcome.

# Please note when copying and pasting the following commands the { signs
# will be replaced by the letter 'f' and the } signs will be replaced by the letter
# 'g'. Unfortunately these will need to be amended before using the codes.

STRATEGY1=function(p,q,h,d,a,r){
values=(prob(p,q))/c(1/h,1/d,1/a)
BET=ifelse(values > r,1,0)
return (BET) }

# For the example in Section 5.7 (Example 3) p = 2.7238587, q = 0.7663088,
# h = 1.15, d = 9.95, a = 33 and we used a value of v = 1.1:

# > STRATEGY1(2.7238587, 0.7663088,1.15, 9.95, 33.00, 1.1)
#     HOME DRAW AWAY
#    [1,] 0 1 1.
```

Appendix E. Betting Strategy for over/under 2.5 goals

```
# Here p and q are the same as previously, o and u should be EU odds for
# over or under 2.5 goals respectively and again r is the discrepency level.

STRATEGY2.5=function(p,q,o,u,r){
values=(prob2.5(p,q))/c(1/o,1/u)
ODDS=cbind(o,u)
BET=ifelse(values>r,1,0)
return (BET) }

# For the example in Section 5.8 (Example 4) p = 0.9350388, q = 1.6870622,
# o = 1.5, u = 3 and again we used a value of v = 1.1:
```

```
# > STRATEGY2(0.9350388, 1.6870622, 1.5, 3, 1.1)
#     over under
#      [1,] 0 1
```

References

[1] Maher, M.J. (1982). Modelling association football scores. Statistica Neerlanda, 36 (3), 109 118.

[2] Dixon, M.J., Coles, S.G. (1997). Modelling association football scores and inefficiencies in the football betting market. Applied Statistics, 46 (2), 265-280.

[3] Pope, P.F., Peel, D.A. (1989). Information, Prices and Efficiency in a Fixed-Odds Betting Market. Economica, 56, 323-341.

[4] Dixon, M.J., Robinson, M.E. (1997). A birth process for association football matches. The Statistician, 47 (3), 523-538.

[5] Cain, M., Law, D., Peel, D.A. (2000). The favourite longshot bias and market efficiency in UK football betting. Scottish Journal of Political Economy, 47 (1), 25-36.

[6] Crowder, M., Dixon, M.J., Ledford, A., Robinson, M. (2002). Dynamic modelling and prediction of English Football league matches for betting. The Statistician, 51 (2), 157-168.

[7] Karlis, D., Ntzoufras, I. (2003). Analysis of Sports Data by Using Bivariate Poisson Models. The Statistician, 52 (3), 381-393.

[8] Hirotsu, N., Wright, M. (2003). An evaluation of characteristics of teams in association football by using a Markov process model. The Statistician, 52 (4), 591-602.

[9] Dixon, M.J., Pope, P.F. (2004). The value of statistical forecasts in the UK association football betting market. International Journal of Forecasting, 20 (4), 697-711.

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

[11] Graham, I., Stott, H. (2008). Predicting bookmaker odds and efficiency for UK football. Applied Economics, 40, 90-109.

[12] Owen, A. (2010). Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. IMA Journal of Management Mathematics, 22, 99-113.

[13] Constantinou, A.C., Fenton, N.E., Neil, M. (2012). Pi-football: A Bayesian network model for forecasting Association Football match outcomes. Knowledge-Based Systems, 36, 322-339.

[14] Koopman, S.J., Lit, R. (2012). A dynamic Poisson Model for Analysing and Forecasting Match Results in the English Premier League. Tinbergen Institute Discussion Paper.

[15] Vlastakis, N. Dotsis, G.Markellos, R.N. (2009). How efficient is the European football betting market? Evidence from arbitrage and trading strategies. Journal of forecasting, 28 (5), 426-444.

[16] Clarke, S.R. Norman, J.M. (1995). Home Ground Advantage of Individual Clubs in English Soccer. The statistician, 44 (4), 509-521.

[17] Hogg, R.V. Introdction to mathematical statistics (7th edn), Pearson.

[18] Haight, F.A. Handbook of the Poisson Distribution, Publications in Operations Research.

[19] Johnson, N.L. Kotz, S. Kemp, A.W. Univariate Discrete Distributions (2nd edn), Wiley series in Probability and Mathematical Statistics.

[20] Sakamoto, Y. Ishiguro, M. Kitagawa, G. Akaike Information Criterion Statistics, D. Reidal Publishing Company.

[21] Everitt, B.S. The Analysis of Contingency Tables (2nd edn), Chapman and Hall.

[22] Football-Data (2012). www.football-data.co.uk. Historical data for English football leagues downloaded from http://www.football-data.co.uk/englandm.php
[23] The Guardian. www.guardian.co.uk. Article on Paddy Power Profits can be seen at http://www.guardian.co.uk/business/2013/mar/06/paddy-power-reports-surge-profits
[24] Download Index for linux, macos, macosx and windows. Software download available free at http:/http://cran.r-project.org/bin/
[25] Online betting website. http://www.online-betting.me.uk. Odds definitions found at http://www.online-betting.me.uk/articles/understanding-betting-odds.html
[26] Odds Portal website. http://www.oddsportal.com. Sure-bets available at http://www.oddsportal.com/sure-bets/
[27] Bet brain website. http://www.betbrain.co..uk. Value-bets available at http://www.betbrain.co.uk/valuebets/
[28] William Hill website. sports.williamhill.com. Terms and conditions can be found at the bottom of the page of sports.williamhill.com.
[29] Github website. https://github.com/maunderb/Betting-Strategy. This is a direct link to the website that holds the R codes for this report.