Robust Financial Prediction by Learning from the Collective Intelligence of Experts

Nicolai Skov Johnsen

Kongens Lyngby 2013
DTU Compute-M.Sc.-2013-54

Supervisor: Ole Winther
Co-supervisor: Sune Lehmann Jørgensen

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Building 303B, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253031, Fax +45 45882673
compute@compute.dtu.dk
www.compute.dtu.dk DTU Compute-M.Sc.-2013-54

Abstract (English)

Odds issued by bookmakers may contain generic biases enforced by typical gambling behaviour, which lead to market inefficient odds. By employing a comprehensive dataset of odds from up to 51 bookmakers on the English Premier League and the Spanish La Liga, seasons 00/01-12/13, the existences of such biases are demonstrated. The biases are particularly prominent in the La Liga, suggesting a more irrational betting behaviour.

A theoretical analysis of odds setting techniques reveals that market inefficiencies may also originate from bookmakers' inherent objective to balance their books. A neural network classifier, which applies the odds as input features, has been combined with a decision framework based on optimization of the standardized expected return per match to profit on the inefficiencies. Two modifications of the betting model have been proposed. Firstly to accommodate a model bias to engage odds selections with overestimated posteriors, secondly to restrict the model to certain probabilistic regions, in which the odds segment evidently is more profitable. It has been demonstrated that the model has high probabilistic accuracy and profits significantly on the La Liga, although the returns are generally season dependent. With the inclusion of the posterior restrictions the model yields the highest and most robust annual return of 16% on the La Liga. The neural network's predictive accuracy is indifferent to whether 5, 9 or 37 bookmakers' odds are used as input features, indicating a low data complexity. Unsolved issues remain regarding the selection bias and refinements of the probabilistic restrictive model.

Abstract (Danish)

Odds udstedt af bookmakere kan indeholde generiske bias, der er frembragt af almindelig spilleadfærd. Disse bias kan lede til markedsineffektive odds. Eksistensen af disse bias påvises ved anvendelse af et omfattende datasæt, bestående af odds fra op til 51 bookmakere på den engelske Premier League og den spanske La Liga, sæson 00/01-12/13. Disse bias er særligt tydelige i La Liga, hvilket indikerer en mere irrationel spilleadfærd.

En teoretisk analyse af hvordan bookmakere sætter odds påviser at ineffiktive odds også kan udspringe fra bookmakeres grundlæggende målsætning om at balancere odds-sæt. For at profitere på ineffiktiviteterne udvikles en neural netværks classifer, der anvender odds som input features. Netværket kombineres med et beslutningsframework, baseret optimering af den standardiserede forventede gevinst per kamp. Modellen modificeres på to måder, dels for at tage hånd om en generel tendens til at udvælge scenarier med overestimerede klasse-posteriors, dels for at begrænse modellen til kun at udvælge odds, hvis tilhørende posteriors ligger i særligt profitable intervaller. Det demonstreres at modellen har stor sandsynlighedsmæssig præcision og tjener betydeligt på den spanske liga. Dog er fortjenesten generelt sæsonafhængig. Den største gevinst på omkring 16% pr. sæson observeres på den spanske liga ved begrænsing af klasse-posteriors. Det neurale netværks præcision er uafhængigt af, om der anvendes 5, 9 eller 37 bookmakeres odds som input features. Dette indikerer en lav datakompleksitet. Projektet har en række uafklarede emner, omkring tendensen til at overestimere posteriors og om muligheden for at videreudvikle den restriktive model.

Preface

This thesis was prepared at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark (DTU) in fulfilment of the requirements for acquiring an M.Sc. in Mathematical Modelling and Computation. The workload has been prepared for 32.5 ECTS points.

The thesis deals with statistical methods and machine learning techniques to analyse and exploit the market inefficiencies in the bookmaker industry within association football, primarily focusing on two major European national leagues.

The thesis consists of three main sections. The first section explains preceding efforts within the field of research, and the basic concepts and issues in odds setting. The second section encompasses a statistical analysis of a comprehensive odds dataset, and the third section contains proposals and demonstrations of forecast models to profit on the market inefficiencies. The MATLAB documentation has been uploaded to DTU CampusNet.

Kongens Lyngby, June 27, 2013

Nicolai Skov Johnsen

Acknowledgements

First and foremost I would like to thank my supervisors Associate Professor Ole Winther and Associate Professor Sune Lehmann Jørgensen for their valuable guidance and advice. I would like to thank Ole for his commitment during the weekly meetings and his responsiveness and patience during several email correspondences. I would also like to thank Sune for his creative suggestions to the project and for sharing his comprehension of the social scientific issues related to the topic of the thesis.

Furthermore I would like to thank the authority of the Department of Applied Mathematics and Computer Science at DTU for providing a good working environment during the completion of the thesis, as well as the relevant staff for providing a freely available MATLAB toolbox for neural network classification.

Finally I would like to thank the website providers at `football-data.co.uk` and `betexplorer.com` for their freely accessible historical football odds archives.

Contents

Chapter 1

Introduction

1.1 Background

The bookmaking industry has been a subject to a significant global expansion during the latest two decades. Improved television coverage and, most importantly, the growing Internet accessibility have created a beneficial commercial environment for bookmakers to attract customers to their betting services. While the technological development has led to increased interests from more bookmaking actors with desires to benefit on the new conditions, it has also enhanced the market competition effectively pushing the odds upwards. The intensification has promoted the need for bookmakers to apply more sophisticated odds setting techniques to optimize their books, as inefficient odds are increasingly financially penalized. Today bookmakers offer a variety of exciting betting options to attract customers, while attempting to conceal the inherent goal of earning money. The greater supply and higher odds have attracted more and more professional gamblers relying on increasingly complex methods to profit on the market. Evidently the battle between bookmakers and gamblers constitutes an ever ongoing arms race. This thesis jumps into the conflict and illuminates the issues related to odds setting with emphasis on association football, from now on referred to simply as football, and propose a forecast model to profit on the market and 'beat the bookies'.

1.2 A Historical Review

Betting on sport events has been dated back to Greece more than two thousand years ago, where wages were made on e.g. the Olympic Games. Gambling was further developed as a business by the Roman Empire, particularly on the gladiator games [1]. In modern history betting on horse racing has traditionally been a part of the British sports culture for centuries.

The UK bookmaker industry has its origins in the 18th century. Initially gambling odds where offered on individuals, typically the favoured horse against the field [2, p. 89], but eventually the betting offerings were expanding by enabling betting against all horses. This formed the basis of the modern *book* [2, p. 90] - i.e. a record of bets placed on a race. Typically the betting contracts were made with no physical money exchange causing severe social damage including large debts and hostility [3]. This lead to the enforcement of the 1845 UK Gaming act, whose main policy was to discourage gambling. The act overturned gambling as legal contract meaning that no gambling debts could be settled by law [2, p. 90]. The act also restricted betting to take place only on race tracks, and so special excursion trains were established to transport people to and from the events. This offer attracted all classes of society, increasing the popularity of bookmaking [4]. Soon many thousands of betting shops began to emerge in the UK. The shops were outlawed by the 1853 Betting Acts but were eventually legalized by the 1960 Betting and Gaming Act [3]. Today the group of the largest bookmaker companies in the UK is known as 'the big three', consisting of William Hill, Ladbrokes, and Coral [5].

Gambling has also developed rapidly in many other countries, as a result of improved TV coverage and modernization of gambling laws. Not least the massively improved Internet accessibility has encouraged many bookmakers to establish on-line brands, often combining traditional sports bookmaking with on-line casino games. The massive explosion of online gambling offers is considered a major cause of the increased gambling addiction in the UK [6]. Today the majority of televised European sport events are sponsored by bookmakers. Particularly association football is heavily sponsored, since football fans constitute a significant fraction of the bookmakers' target group. As a result of the banning of tobacco sponsorships and the growth of the gambling industry many sponsorships of major European football teams are now taken over by bookmakers instead of car manufacturers and soft drinks producers.[7]

For a long time each European country has managed its own gambling legislation. However, the forming of the European Union has caused a lot of uncertainty for the gambling market, as the EU permits open access to all EU countries gambling markets. Many European counties have tried to enforce

their laws prior to protect their markets or to gain advantage over other's markets. [8] In Denmark a liberalization of bookmaking was induced the 1st of January 2012. Until then 'Danske spil', mainly owned by the Danish state [9], had monopoly on bookmaker business in Denmark [10]. In USA the gambling laws are regulated by each territory, but almost all states consider gambling a legal activity. Generally the legislation is relaxed more and more each year [11]. In Asia gambling is illegal in for example China and Japan among others [12].

In the later years an increasing number of gamblers have switched to betting exchanges. A bet exchange essentially provides a trading facility where *punters* (the bettors) can buy or sell gambling contracts to each other. Many consider bet exchanges more attractive than traditional bookmakers, partly because the exchanges do not restrict the size of the bets, which are only limited by the willingness of opposing costumers [13]. Generally exchanges also provide better odds. Whereas traditional bookmakers earn their money on *overrounds*, the exchanges charge a commission on winnings instead. This effectively reduces the bookmakers' percentage profit margins on their books, the so-called overrounds, to zero [14].

1.3 Forecast Methodologies

Publications regarding statistical methods for football predictions have predominantly been appearing over the latest two decades. However, the earliest work goes back to 50s-70s in which researchers primarily focused on modelling the distribution of the goals scored in a match and not on the profitability of such models. Not until the 90s, a combination of forecast models and betting strategies were proposed to detect and utilize the inefficiencies in bookmakers' odds. The inefficiencies consist of selections with disproportionally high odds compared to the success probability. In this section some of the major statistical techniques are outlined, in the field of football match prediction.

1.3.1 Negative binomial distribution goal-based analysis

The first statistical analysis of football results was released by [15] (1951) focusing on efficiently modelling the number of goals scores by single teams in football matches. The author demonstrated that the number of goals was poorly fitted by the family of Poisson distributions and found that a negative binomial distribution provided a far more adequate fit to the observations.[16]

The conclusions were confirmed by [17] (1971) on a variety of ball games. It was perceived by the authors that the negative binomial distribution can be successfully applied to the goal score, provided that the chance of winning is invariant to the strength of the opposing teams. In situations where the individual team skills played a stronger role the model had poor forecast capability, and the authors concluded that 'chance does dominate the game'.

1.3.2 Poisson distribution goal-based analysis

In contrast to [17] it was demonstrated by [18] (1974) that football experts were indeed able to predict the final league positions to a certain extent, indicating that the strength of the teams dominates the outcome and not simply chance. Presumably motivated by this conclusion [16] (1982) proposed the first model to predict the outcomes of football matches in games accounting the individual team strengths. The underlying assumption of the model was that ball possession is essential to the number of goals scored by the opposing teams. Although earlier researchers concluded that the Poisson distribution was insufficient to model the goal score Ref. [16] recommenced the approach by proposing an independent Poisson model. If the home team i is playing against team j, the observed score can be formulated as (x_{ij}, y_{ij}), where the goal scores x_{ij} and y_{ij} follow two independent Poisson distributions with means $\alpha_i\beta_j$ and $\gamma_i\delta_j$, respectively. In this configuration α_i and γ_i represent the strength and weakness of team i when playing home, and β_i and δ_i represent the strength and weakness team j when playing away. Hence if 22 teams are playing in a league, a total of 88 parameters is required. However by enforcing the constraints $\sum_i \alpha_i = \sum_i \beta_i$ and $\sum_i \gamma_i = \sum_i \delta_i$ only 86 independent parameters need to be uniquely determined. The author used data from Rothmans Football Yearbook (1973-1975), which contains results from British football leagues. Maximum likelihood estimates of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ revealed that an adequate simplification only required the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, as all teams were found to be equally affected by the significant home team advantage, and thus the scoring power of each team was diminished by a constant scaling factor K when playing away. It was demonstrated that the model gave a reasonably good fit to the data by comparing the expected and observed score distributions in a χ^2-test. Although the author expressed great content with how well the simple model explained the data, he also stressed the possible insufficiencies. 'A match does not consist of two independent games at opposite ends of the pitch', he remarked. Consequently, an bivariate Poisson model was proposed in which the marginal distributions were still Poisson with means $\alpha_i\beta_j$ and $K\alpha_j\beta_i$ but with an additional correlation ρ between the scores. By varying ρ it was found that $\rho = 0.2$ appeared to be most appropriate. It was demonstrated that this model yielded a considerably better fit on the differences in scores.

The basic model structure in [16] was refined by [19] (1997), as the authors proposed a set of modifications to evade some of the limitations of the original model. Unlike preceding publications the primary motivation of [19] was to profit on the inefficiencies in the bookmakers' odds. It was demonstrated that independence between scores is a reasonable assumption except for low scoring games: 1-1, 1-0, 0-1, 0-0 and that the bivariate Poisson distribution family, as proposed by [16], was unable to account for the *varying* dependency. Accordingly, the independent Poisson model was modified as follows

$$P(X_{ij} = x, Y_{ij} = y) = \tau_{\lambda,\mu}(x, y)Pr(x; \lambda)Pr(y; \mu),$$

where $\lambda = \alpha_i\beta_j$ and $\mu = K\alpha_j\beta_i$. If $x \leq 1$ and $y \leq 1$ the dependency is perturbed, otherwise the scores are independent, i.e. $\tau_{\lambda,\mu}(x, y) = 1$. Additionally the authors discarded the static strength parameters α_i and β_i and enforced dynamic strength parameters instead. The expected return from a unit stake was formulated as

$$E(\pi_p) = \frac{p}{b} - 1,$$

where π_p is the punter's return, p is estimated probability on the engaged selection and b is probability on the selection implied by the odds as the normalized reciprocal value of the odds. The authors found a reasonable agreement between p and b and demonstrated that the model yielded a positive return provided that $\frac{p}{b} > r$ for any $r > 1.1$.

1.3.3 Probit regression

Unlike the Poisson models all other proposed models restrict the analysis to forecast match results. Models based on probit regression are among these. Ref. [20] (2000) proposed a model based on an ordered probit function which basically is a generalization of the probit function to solve multi-class classification problems. The author formulated a model to illuminate bookmakers' odds setting decisions and found that odds market inefficiencies are possible if bookmakers strive to maximize their expected profit. Contrary, if the bookmakers use a risk-minimizing strategy whereby the odds agree with the subjective probabilities derived by the bookmakers, the market would be efficient. The probit model was evaluated using different combinations of odds and publicly available statistics as explanatory variables. The statistics comprised of a set of differences in records between the opposing teams such as points and league positions. An empirical test on English football leagues with the same decision criterion as stated in Eq. 1.3.2 proved that market inefficiencies indeed are prevalent. The author suspected that the inefficiencies are consequences of team loyalty which the bookmakers take advantage of by setting market inefficient odds.

Similar studies of probit forecast models that applies published bookmakers' odds are found in [21] (2005), where the forecast effectiveness of a probit model based on publicly available statistical data is evaluated and compared to the probabilities implied by the odds. The model was applied on a dataset consisting of data from British bookmaking firms during five seasons 98/99-02/03. The authors observed that the probabilistic model was superior in the early seasons. However, in the later seasons the opposite was true, suggesting an improved expertise of bookmakers as forecasters, and that bookmakers utilize information not included in the public statistics. It was perceived that such tendency is a consequence of increased competition in the bookmaker industry, whereby inaccurate odds setting is increasingly penalized from a financial perspective.

1.3.4 Bayesian networks

Bayesian networks (BN) are probabilistic models that apply a graphical approach to explain the conditional dependencies of random variables through a system of directed acyclic graphs. Essentially each node in the graph corresponds to a random variable and the links express the probabilistic relations between the nodes. [22, pp. 359-662] A Bayesian network can be utilized to solve decision problems by estimating the probabilities of different events.

In recent years researchers have showed a high potential of BNs which consider historical data as well as expert judgements to forecast football match results. Specifically [23] (2006) applied an expert constructed BN to forecast matches involving the Tottenham Hotspurt. The model considered a collection of four features: 1) A binary variable stating the presence or absence of three key players on Tottenham Hotspurt; 2) the field position of a key player; 3) expert judgement of the quality of the opposing team; and 4) whether the match was played on home or away field. The authors demonstrated that the expert-based BN had higher forecast accuracy than four alternative machine learning models: Naive Bayes, Bayesian learning K-nearest neighbour and decision tree.

In [24] (2012) the authors proposed an expert constructed BN to predict the outcome of matches in the English Premier League season 10/11, using objective variables captured by historic statistics as well as subjective variables. The model considered four generic components on both the home and away teams: 1) strength; 2) form; 3) psychology and; 4) fatigue. Contrary to preceding approaches the authors replaced each team name in each match by a predetermined team strength distribution derived from the total number of points the particular team had archived during the considered season. The two opposing teams' strength distributions were compared which generated an objective forecast. The latter three components were predominantly derived by subjective

information and were used to revise the objective forecast. The authors proposed a proximity scale from 0 to 1 on each component with 0.5 meaning no advantage to either of the teams. Using a standard (unspecified) profitability measure with a fixed discrepancy level as betting rule the overall profit/loss ratio was measured in the Premier League season 10/11. In general the model performed poorly on low discrepancy levels (1%-3%) and much better on higher discrepancy levels (4% - 11%). The model was tested using the maximum (best) odds available, the mean odds and the odds from a single major UK bookmaker. Table 1.1 summarizes the ratios and the percentages of engaged matches with different discrepancies when applying the best odds. Levels exceeding 11% was considered to imply too few bet instances to derive meaningful conclusions. Considering both accuracy and profitability measures the authors demonstrated the significance of the subjective components on the model's forecast capability.

Discrepancy [%]	Profit/loss [%]	Bet fraction [%]
≥ 5	8.40	44.5
≥ 6	13.3	34.5
≥ 7	12.1	28.2
≥ 8	10.0	22.1
≥ 9	16.0	18.7
≥ 10	20.4	13.7

Table 1.1: Profitability statistics on different discrepancies on an expert constructed Bayesian network by Ref. [24] on Premier League season 10/11.

1.3.5 Neural networks

Artificial neural network (NN) models for classification constitute a highly flexible class of probabilistic models. A further description of NNs is found in section 4.1. Ref. [25] (2011) proposed a simple NN model to estimate the probability of the possible outcomes which solely considered bookmakers' odds from a variety of bookmakers as input features. Using an ensemble of NNs a Dirichlet distribution was fitted to the estimated outcome probabilities. This enabled a further analysis of the probability of a value bet, that is, a bet with positive expected return, and the probability of winning on bets. The model was evaluated 5 runs on the Spanish La Liga seasons 07/08-10/11. Each season contains 380 matches and so the total match count is $380 \cdot 4 = 1520$. The test partition consisted of 400 random samples and the training partition contained the remaining samples. The simulations yielded an average profit per unit bet of 0.16.

1.4 Common Betting Systems

This section outlines two of the most basic strategies which forex traders among others use to bet on the financial market [26] in an attempt increase the chance of a long term profit. Although extensive literature exists on the mathematical background of these strategies this is only briefly described. Focus will be placed on illuminating the concepts and the prevailing psychological biases that affect the behaviour of many gamblers.

1.4.1 The martingale system

A martingale system is a betting management system where the investment continuously increases after each loss on a betting market. The most simple and possibly best known example of a martingale system is a coin flip, where the actor wins her/his stake should the coin come up head and loses the stake otherwise. Whenever a loss occurs the stake is doubled in anticipation of the statistical guarantee of a future increase in fortune. Although it is perceived by many as a lucrative strategy it is simply an example of a cognitive bias to neglect the importance of low probability events, such as a series of losses. The martingale strategy only proves itself as a winning strategy if the gambler has infinite wealth and time, and if the market provider has no limits on the bet sizes. These assumptions limit the application of the strategy.

The coin flip is a fair game, i.e. a game where the expected return on each flip is equal the last observation. This is an example of a *martingale* sequence, defined as a stochastic process in which a sequence (X_n) of random variables at any discrete time $n \in \mathbb{N}$ satisfies

$$E[X_n] < \infty \tag{1.1}$$

$$E[X_{n+1} \mid X_1, \ldots, X_n] = X_n \tag{1.2}$$

From the properties of conditional expectation it is seen that $E[X_{n+1}] = E[X_n]$ and accordingly $E[X_n] = c$, $\forall n \in \mathbb{N}$ for some constant c [27, p. 150]. Hence the expected net return of the coin flip is in fact zero, since $E[X_n] = 0$. The concept of martingales can generalized to cases, where the expected value of future observations need not be equal the last observation known as sub-martingales where $E[X_{n+1} \mid X_1, \ldots, X_n] \geq X_n, \forall n \in \mathbb{N}$ and super-martingales where $E[X_{n+1} \mid X_1, \ldots, X_n] \leq X_n, \forall n \in \mathbb{N}$. The latter sequence constitutes a game series that is unfavourable to the gambler [28, p. 96] which represents set-up in casinos etc. Essentially you cannot beat the system, since the odds are never fair to the gambler. The former sequence is unfavourable to the book-maker. A sequence of bets with positive expected return would ideally guarantee

a positive net profit. However, the inherent limitation of wealth makes the application questionable.

1.4.2 The anti-martingale system

Contrary to the martingale system the anti-martingale system urges the gambler to increase bets after a win and reduce bets after a loss [29]. The approach is based on the perception of that the gambler can benefit from a sequence of wins, while reducing losses on a sequence of losses. However, if single bets are independent of each other the notion of such streaks is absurd and is simply an example of the Monte Carlo fallacy. It basically means that the gambler is biased to believe that the probability of outcome O after a series of outcomes O is less than before the series [30]. On the other hand, if the bets are serially correlated, one could benefit from the strategy. Many on-line traders consider the anti-martingale less risky than the conventional martingale system, since it is perceived to be less risky to increase trade sizes during a streak of wins than during a streak of losses. [26] Presumably the trading system constitutes a dynamic system, where economic booms and slumps appear in cycles, whereby the trades indeed are correlated.

1.5 Markowitz Portfolio Optimization

A far more general concept in finance and a cornerstone in modern portfolio theory is the so-called efficient frontier. Given a desired expected return R, the efficient frontier represents the minimum risk portfolio. In portfolio theory the risk corresponds to the standard deviation σ of the portfolio's return, and hence the efficient frontier essentially defines the line in the risk-return space (σ, R), which optimizes the trade-off between risk and expected return.

Consider a portfolio of N assets with return $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{C})$, $\boldsymbol{y}, \boldsymbol{\mu} \in \mathbb{R}^N$, $\boldsymbol{C} \in \mathbb{R}^{N \times N}$ where y_i and μ_i denote the return and expected return, respectively, of the ith asset and $\boldsymbol{C}$ holds the covariance among the assets. Furthermore denote the weight of the assets $\boldsymbol{w} \in \mathbb{R}^N$ with the constraint $\sum_{i=1}^{N} w_i = 1$, where w_i is the weight of the ith asset. The equations that govern the efficient frontier can

then be stated as a convex quadratic optimization problem

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w} \\ \text{subject to} \quad & \boldsymbol{w}^T \boldsymbol{\mu} = R, \quad \sum_{i=1}^{N} w_i = 1 \end{aligned} \tag{1.3}$$

where $\boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w}$ is the portfolio variance. Accordingly, for a given desired expected return R, there is an optimal portfolio characterized by $\boldsymbol{w}$ with a corresponding point on the efficient frontier (σ, R). The line of optimal weights is commonly referred to as the *minimum-risk weight line*. The problem is convex since $\boldsymbol{C}$ is positive semi-definite, and thus the minimum-risk weight line forms a convex curve in the (R, σ) space, commonly known as the *Markovitz bullet*. [31, pp. 41-60]

1.6 Thesis Statement

The purpose of this thesis is to perform a statistical analysis of a dataset consisting of full time result odds from association football matches and to develop and test a probabilistic betting model on the dataset.

The dataset will be obtained by implementing a MATLAB API to extract free on-line statistics on odds from 51 bookmakers covering the English Premier League and the Spanish La Liga, seasons 00/01-12/13. The data will be used to investigate the extension of cognitive biases in gambling behaviour and odds setting, and if there is evidence of a non-stationary odds market with increased market competition. Additionally the data will be utilized to derive statistical characteristics of odds in terms of match outcomes and to examine to what extend bookmakers differentiate in odds setting.

Based on the odds analysis a betting model will be proposed, consisting of a neural network classifier combined with a decision framework, to address the issues of when and how much to bet to obtain a long term profit. The quality of the model will be evaluated in terms of probabilistic accuracy and profitability, and the model will be used to elaborate the prospective biases inherited in the odds.

Chapter 2

Concepts and Operational Procedures in Odds Setting

2.1 Odds Representations and Common Bet Types

The most prevalent type of odds in bookmaking and the only type processed in this thesis is called *fixed odds*. Fixed odds refer to a market where the bookmaker and punter sign a contract on a fixed rate of *return* and on a fixed amount staked on a given *selection*. The *return* or *gain* refers to the total amount of money gained by the punter if she/he wins, and the selection refers to the result of an event that the punter bets on.

The most widely used meaning of quoted odds (except in USA) are *decimal odds* and *fractional odds*. The decimal odds notation is favoured in Europe, and corresponds to the factor by which the *punter* (bettor) may multiply his/her bet which yields the return if the punter wins on his/her selection. The examples and mathematical notation in this thesis are solely based on the decimal odds notation.

The *fractional odds* notation is preferred in the UK. A fractional odds is often noted as x/y, $x-y$ or $x:y$, $x, y \in \mathbb{N}^+$. A chance of e.g. 2-to-9, meaning a fractional odds of $x/y = 2/9$, is equivalent to a decimal odds value of $\frac{x}{y} + 1 =$

$\frac{11}{9}$. Hence the fractional odds describe the relative occurrences of winnings compared to all other alternatives. A special case is the 1/1 fraction which is often shortened to an 'evens' and quoted as 2.0 in decimal odds. [32] The fractional odds notation will not be used in this thesis.

The bookmaker market comprises a variety of different odds configurations. The next sections describe some of the most common betting strategies in football, where the selections encapsulate the full time results. That is, only odds for home win, draw and away win are considered. The reader should be aware that there are many other popular types of bets in football such as the first goal scorer, the correct full time score, the first yellow card in the match, etc. [33]. Another popular strategy is the *each-way bet*, consisting of a *win bet* and a *place bet* of equal size with a major and minor return, respectively. In football each-way bets are typically used in e.g. the finishing positions of teams or the top score list in domestic leagues. The win bet gives a return if the selection wins and the place bet gives a return if the selection either wins or obtain a second place [34]. All these betting strategies are omitted in the thesis. For convenience the following useful notation is introduced:

Let $\tau_m B$ denote the return of the punter (incl. bets), where $B > 0$ is the amount waged and $\tau_m > 0$ is the overall *odds multiplier*. τ_m is the combined factor of all decimal odds affecting the bet.

2.1.1 Multiple bets

The full time result bet is a member of a larger class of bets known as *single bets*, which is the simplest betting strategy as money is waged on a single event. Another widely popular class of bets is *multiple bets*, where money is waged simultaneously on different events. Only if all selections win, the punter receives a return. Suppose $o_A, o_B, o_C, \ldots$ are the odds then $\tau_m = \Pi_{i=\{A,B,C,D,\ldots\}} o_i$. Specifically if two or three selections are engaged, the bet is called a double or treble, respectively. If more than three selections are engaged it is known as an accumulator.[35]

While multiple bets generally offer high returns, the chances of winning are relatively small, since all selections must win. An alternative family of bets where not all selections need to win, is known as *full cover bets* as discussed in the next section.

2.1.2 Full cover bets

The main idea of full cover bets is to offer a betting strategy, where all selections do not need to win to ensure a return. While there are many full-cover configurations, the general idea is to cover a large amount of the possible combinations of single bets and multiple bets on $N \geq 3$ selections. The greater coverage compared to multiple bets makes it a very popular betting strategy. The full cover bets can be divided into two subclasses, where one subclass contains the singles and the other does not.

2.1.2.1 Full cover bets with singles

In all full cover bets with singles the punter is guaranteed a return if at least one of the selections wins. A *patent* is the smallest member of the family consisting of 3 selections with

$$\tau_m = (o_A + 1)(o_B + 1)(o_C + 1) - 1 \tag{2.1}$$
$$= o_A o_B o_C + o_A o_B + o_B o_C + o_A o_C + o_A + o_B + o_C \tag{2.2}$$

The bet consists of 1 treble, 3 doubles and 3 singles yielding a total of 7 selections. Hence a 1 coin patent bet actually needs a stake of 7 coins. Suppose only two selections win (e.g. A and B), then the odds multiplier is reduced to $\tau_m = o_A o_B + o_A + o_B$. If only one selection wins (e.g. A), then $\tau_m = o_A$. So one selection is enough to guarantee a return.

Full cover bets with singles with $N = 4, 5, 6$ are known as *Lucky 15*, *Lucky 31* and *Lucky 63*, respectively. These bets follow the same principles as the patent. Lucky 15 consists of 1 fourfold accumulator, 4 trebles, 6 doubles and 4 singles which gives a total of 15 selections, i.e. $\tau_m = \Pi_{i=\{A,B,C,D\}}(o_i + 1) - 1$. Hence the name Lucky 15. Similarly Lucky 31 and Lucky 63 have 31 and 61 selections, respectively. [35]

2.1.2.2 Full cover bets without singles

In all full cover bets with singles, the punter is guaranteed a return if at least two of the selections wins. The smallest member of this subclass is known as a *trixie*, consisting of three selections. Using the same notation as in subsection 2.1.2.1, τ_m can be expressed as

$$\tau_m = (o_A + 1)(o_B + 1)(o_C + 1) - o_A - o_B - o_C - 1 \tag{2.3}$$
$$= o_A o_B o_C + o_A o_B + o_B o_C + o_A o_C \tag{2.4}$$

The bet consists of 1 treble and 3 doubles, i.e. a total 4 selections. Therefore a 1 coin trixie bet requires a stake of 4 coins. In case only two selections win, τ_m will only be equal the winning double. Full cover bets without singles with $N = 4, 5, 6, 7, 8$ are known as *Yankee*, *Super Yankee*, *Henz*, *Super Henz* and *Goliath*, respectively. [35]

Compared to the full cover bets with singles, the full cover bets without singles have less coverage, but require a smaller stake for the same number of selections. The largest contributions to τ_m is the highest order multiple bets, so if a punter is mainly interested in these, the full cover bets without singles might be preferred.

2.2 The Concept of Overround

A bookmaker's long term profit, denoted π, depends the the so-called *overround*. This is the amount by which a *book* exceeds 100%, corresponding to a profit margin. A book is a set of odds covering different outcomes of an event. To put it more formally:

Consider a sport event with a book consisting of N odds, denoted o_i, $i = 1, \ldots, N$. The overround, denoted by η, is then found as

$$\eta = \sum_{i=1}^{N} o_i^{-1} - 1 \tag{2.5}$$

where o_i^{-1} can be interpreted as the relative probability of the i'th result of the event. The overround η serves as a general measure of the bookmaker's margin of safety.

In general bookmakers have high profits in markets with many possible outcomes on single events. The more possible outcomes the less probable the individual outcomes will be [36]. Suppose e.g. that an event comprises N selections on which a given bookmaker perceives equal chances of successes $p = \frac{1}{N}$. Accordingly, the odds o_i, $i = 1, \ldots, N$ are set equally as $o_i = o^* = \frac{1}{p(\eta+1)} = \frac{N}{\eta+1}$, where η is the bookmaker's profit margin. Without loss of generality assume that a unit bet is made on each selection. The bookmaker's profit is then $\pi = N - \frac{N}{\eta+1} = \left(1 - \frac{1}{\eta+1}\right) N$. Provided that $\eta > 0$ the profit increases proportionally with N.

Another highly lucrative configuration for bookmakers is *multiple bets*, see section 2.1, where the punter bets on a series of events. The overall overround η' on

multiple selections increases as the overround η on each event is compounded. To formalize this insight suppose e.g. that a punter chooses to bet on a series of two events provided by a given bookmaker, each with N selections. For simplicity assume that the overround η is the same on both events. Further let o_{ij} denote the odds on the j'th selection in the ith match, which is found as

$$o_{ij} = \frac{1}{p_{ij}(\eta + 1)}, \quad i = 1, 2, \quad j = 1, \ldots, N \tag{2.6}$$

If the punter chooses single selections m and n in each match, the total prize on a unit bet if the punter succeed in both selections is $o_{1m}o_{2n} = \frac{1}{p_m p_n (1+\eta)^2}$. Thus the overall overround is $\eta' = (1+\eta)^2 - 1 = \eta^2 + 2\eta$. Provided that $\eta > 0$ it is seen than $\eta' > \eta$. The bookmakers' high margins of safety on multiple bets are unattractive to the punters.

In a *fair bet* the overround is zero, but such a zero-sum configuration is obviously undesirable for any bookmaker. The higher the overround the more the bookmaker should statistically earn. However, strong competition in the bookmaker industry drives the odds up and the overround down.

2.3 Modelling the Odds Setting

The main objective of any bookmaker is to guarantee profit by archiving what is commonly known as a *balanced book*. This risk management procedure essentially means that the bookmaker strives to obtain a bet distribution so that the bookmaker profits equally on an event regardless of the outcome. Bookmakers do not profit from the bets themselves but operates as a market makers. That is, bookmakers offer odds at a price that is higher than the expected payout to the punters, cf. section 2.2. The concepts are similar to actuary in e.g. the insurance industry, where one attempts to balance the financial outcome of events. [37]

When setting the odds values in a book the bookmakers subjectively estimate the probabilities of the outcomes from which they set the odds. However, no bookmaker wants to render its intentions too visible. Instead the odds are set at what is perceived by public to correspond to the 'true' probabilities in order to reduce the imbalance of the book. The potential *market inefficiency* enforced by the public opinion is important in order for punters to make a long term profit. In the remainder of the thesis only full time result (FTR) odds on football matches are considered, that is the home, draw and away odds, denoted o_1, o_2 and o_3, respectively.

2.3.1 Ideal balancing of books

Suppose that a bookmaker offers a book on the market and that the odds are completely fixed. The following definition states the ideally required distribution of bets on each selection in order to obtain a perfectly balanced book.

Consider a football match and let o_i, $i = 1, 2, 3$ denote the odds on the ith selection (home, draw and away) for a given bookmaker X. Assume that the total money waged on X is $B = b_1 + b_2 + b_3$, where b_i is the bet on the i'th result. Given B and o_i, $i = 1, 2, 3$ so that $\eta > 0$, the bets b_i, $i = 1, 2, 3$ are uniquely determined in order to ensure that the bet is balanced. (See proof in app. A.1.) The value of these unique bets are given by (see proof in app. A.2)

$$\boldsymbol{b} = [b_1, b_2, b_3]^T = \left[\frac{\pi}{\eta o_1}, \frac{\pi}{\eta o_2}, \frac{\pi}{\eta o_3}\right]^T \tag{2.7}$$

where bookmaker X's profit, regardless of the outcome, π is given by

$$\pi = B\frac{\eta}{\eta + 1} \tag{2.8}$$

Hence, for every $1+\eta$ coins betted, the bookmaker profits η coins. The fraction $\frac{\eta}{\eta+1}$ is therefore the bookmaker's relative profit on the book if the book is balanced. The example below demonstrates this result.

Example: Balancing odds Consider a football match with odds

$$\boldsymbol{o} = [o_1, o_2, o_3]^T = [1.25, 5.80, 7.25]^T$$

The overround is then

$$\eta = \sum_{i=1}^{3} o_i^{-1} - 1 \approx 0.1103 \tag{2.9}$$

The bookmaker produces an equal return, if and only if the waged money on the selections is ideally proportioned to the odds, c.f. the definition above. Assuming a total amount bet of $B = 100(1+\eta) \approx 111.03$ the amount betted on each selection to ensure a balanced bet is found as, cf. Eq. 2.7

$$\boldsymbol{b} \approx [80.00, 17.25, 13.79]^T \tag{2.10}$$

with the return equals, cf. Eq. (2.8)

$$\pi \approx 111.03 \cdot \frac{0.1103}{1.1103} = 11.03 \tag{2.11}$$

I.e. for every 111.03 coins betted in a balanced bet the bookmaker gains $\pi \approx 11.03$, since $\eta \approx 0.1103$.

2.3.2 The dynamics of book balancing

Although all bookmakers' main goal is to balance their books, a perfect balance is hardly achievable, due to the dynamics in the bookmaker market. Typically the odds are adjusted all the way up to the match, as the bookmakers persistently attempt to optimize the odds setting when new bets are made. The simplifying examples below provide insight into how these mechanisms interact. [1]

Example: Balancing odds in a progressive market Consider a bookmaker X engaging in a football match with the same odds as in (2.9), yielding $\eta = 0.1103$. Suppose the following amounts are waged $B = 503.12$. Assume that the bet initially is perfectly balanced, cf. Eq. (2.7) and Eq. (2.8)

$$\boldsymbol{b} = [b_1, b_2, b_3]^T = [362.5, 78.12, 62.5]^T \tag{2.12}$$

$$\pi = 503.13 \cdot \frac{\eta}{\eta + 1} \approx 50 \tag{2.13}$$

Hence the bookmaker has a perfectly balanced bet with a relative profit of

$$\pi_{rel} = \frac{\pi}{B} = \frac{\eta}{\eta + 1} = 0.0994 \tag{2.14}$$

Obviously this fraction is smaller than η, since

$$\eta = \frac{\pi}{B - \pi} > \frac{\pi}{B} = \pi_{rel} \tag{2.15}$$

Normally early released odds are very conservative in the sense that η is very large, since the strength of the opposing teams is less clear, than just before kick-off [36]. Suppose now that a minor injury has been reported on one of the key players on the away team, implying that $\tilde{b}_1 = 300$ more is bet on the home team. The bookmaker's profit π_i on the ith match outcome is then

$$\pi_1 = b_2 + b_3 - (b_1 + \tilde{b}_1)(o_1 - 1) = -25 \tag{2.16}$$

$$\pi_2 = (b_1 + \tilde{b}_1) + b_3 - b_2(o_2 - 1) = 350 \tag{2.17}$$

$$\pi_3 = (b_1 + \tilde{b}_1) + b_2 - b_3(o_3 - 1) = 350 \tag{2.18}$$

[1] The examples have been adapted from [36].

The bet is now unbalanced to the extent that the bookmaker losses money on home win, and the bookmaker could strive at re-balance the odds. To do so it might adjust the odds to $\tilde{\boldsymbol{o}} = [1.2, 6.6, 7.95]^T$, effectively increasing the attractiveness on the draw and away odds. In this situation many punters make their long term profit, since the increment in draw and away odds are possibly better odds valued than it should compared to the true probability of these outcomes [36]. Note than the overround almost remains unchanged, $\eta = \sum_i \tilde{o}_i^{-1} - 1 \approx 0.1106$. Suppose now that the odds adjustments leads to that additionally $\tilde{b}_2 = 56.82$ and $\tilde{b}_3 = 47.17$ are betted on draw and away odds. The bet is now re-balanced with new gains equal

$$\tilde{\pi}_1 = b_2 + \tilde{b}_2 + b_3 + \tilde{b}_3 - (b_1 + \tilde{b}_1)(o_1 - 1) = 78.99 \tag{2.19}$$

$$\tilde{\pi}_2 = b_1 + \tilde{b}_1 + b_3 + \tilde{b}_3 - b_2(o_2 - 1) - \tilde{b}_2(\tilde{o}_2 - 1) = 78.99 \tag{2.20}$$

$$\tilde{\pi}_3 = b_1 + \tilde{b}_1 + b_2 + \tilde{b}_2 - b_3(o_3 - 1) - \tilde{b}_3(\tilde{o}_3 - 1) = 78.99 \tag{2.21}$$

The relative profit has now declined to

$$\tilde{\pi}_{rel} = \frac{\tilde{\pi}_i}{B} = 0.0871, \quad i = 1, 2, 3 \tag{2.22}$$

This example emphasizes that if a book is imbalanced, and adjustments must be made to re-balance, it may reduce π_{rel}. It is of obvious and strong interest that bookmakers initially model an accurate punter reaction function to optimize the expected profit. Too many adjustments possibly reduces the bookmakers' gains, and so bookmakers must find an equilibrium between the degree of adjustments and the degree of bet balance. Bookmakers are generally not interested in the gain on single events, but on the long run, so the bets are never perfectly balanced. These adjustments are of great importance to punters seeking a long term profit, when the odds are adjusted higher values, than the 'true' probability represents.

Example: Lay-off odds (Continuing the preceding example.) The bookmakers also have the option to 'lay-off' some incoming bets on a selection s with high potential liability on bet exchanges or other bookmakers, offering better odds at $s = 1, 2, 3$. Suppose that e.g. no additional bets are made on bookmaker X after the adjustment of the odds to $\tilde{\boldsymbol{o}}$. To eliminate the liability, should the home team win X might re-invest in another bookmaker Y or bet exchange Y, offering a slightly higher home odds $\hat{o}_1 = 1.22$. If X wages $\hat{b}_1 = 307.38$ coins on home wins at Y it will gain

$$\pi_1 = b_2 + b_3 + \hat{b}_1(\hat{o}_1 - 1) - \tilde{b}_1(o_1 - 1) \approx 42.62 \tag{2.23}$$

$$\pi_1 = b_1 + b_3 - b_2(o_2 - 1) - \hat{b}_1 \approx 42.62 \tag{2.24}$$

$$\pi_1 = b_1 + b_2 - b_3(o_3 - 1) - \hat{b}_1 \approx 42.62 \tag{2.25}$$

Hence the bet has been re-balanced with a profit of $\pi_{rel} = \frac{\pi_i}{B} = 0.0847$, $i = 1, 2, 3$.

Evidently the bookmaker market constitutes a very complex system. The bookmakers not only strive to offer the most attractive odds but also wage on each other if the differences in odds are attractive. This effectively means that the final odds may deviate significantly from the 'ideal' odds, reflecting the bookmakers' subjective probabilities. Hence the implied probabilities from the odds may deviate strongly from the 'true' selection probabilities.

2.3.3 The favourite/long-shot bias

A prevalent phenomenon in virtually all betting markets is the favourite/long-shot bias which bookmakers implement into the odds settings to balance the bets. The bias is motivated by the fact that punters are generally biased to overestimate the chances of the weak team in a match with a very strong favourite. This is an example of risk seeking behaviour, as the punters aim at the selection with highest potential return although the expected return on both the favourite and weak team selections may be indifferent. While this is very profitable to the bookmaker if the favourite win it also exposes the bookmaker to a very high liability should the weak team win, since the odds on the weak team are very high. To reduce the vulnerability the bookmakers tend to reduce the odds on the weak team and increase odds on the favourite team. This will drag more punters to the favourite team thereby balancing the odds [36]. The bias can be regarded as a special feature, which bookmakers take into account when balancing their books. The existence of such discrepancy between the bookmakers' and the public's perception of the outcome probabilities and accordingly the appropriate odds creates an apparent *market inefficiency*.

Example: Adjusting odds to bias At a given football match, suppose bookmaker X estimates the probability of a home team win at 85%, draw at 10% and away team win at 5%. On this basis X should set the odds around $\frac{1}{0.85} \approx 1.18$, $\frac{1}{0.10} = 10$ and $\frac{1}{0.05} = 20$. Adding an appropriate overround the odds is set at e.g. $\boldsymbol{o} = [1.15, 7.50, 15]^T$, yielding $\eta = \sum_{i=1}^{3} o_i^{-1} - 1 \approx 0.070$. To incorporate the bias X might adjust the odds to $\tilde{\boldsymbol{o}} = [1.25, 5.40, 11.90]^T$, still yielding $\eta = \sum_{i=1}^{3} \tilde{o}_i^{-1} - 1 \approx 0.070$.

The bias effectively means that the favourite odds are often more attractive than the long-shot odds. Though the difference sometimes is small it is of great significance to punters, seeking a long term profit [36]. It should however,

be emphasized that the favourite/long-shot bias may the dominated by other dynamic factors related to the temporal development in bet proportions on the different selections as demonstrated in section 2.3.2.

Chapter 3

Data Collection and Analysis

The dataset consists of the full time results (FTR), the FTR odds and the opposing team names from from two major European football leagues: The English Barclay's Premier League and the Spanish La Liga. For convenience the abbreviations PL and LL shall be used for the English and Spanish league respectively. The dataset captures statistics from seasons 00/01-12/13 with varying number of registered bookmakers present in each season and in each match. For the remainder of the thesis the FTR odds will be referred to as simply the odds.

In both leagues 20 national top teams compete each season in a complete combinatorial system, where all teams are playing against one another on home ground and away ground. Denoting $n = 2$ the number of matches between two fixed teams in a season, $t = 20$ the number of teams in a season and $m = 2$ the number of teams in each match each season comprises a total of $n \cdot \binom{t}{m} = 2\binom{20}{2} = 380$ matches. A brief review of each league is provided below.

Barclay's Premier League or simply Premier League is the top of the English professional football league system. The league was founded in 1992 and has hosted 42 different teams. Since the establishment 5 different teams have

won the league, with a superior number of titles to Manchester United (13 titles) followed by Arsenal and Chelsea (3 titles both), and Blackburn Rovers and Manchester City (1 title both).[38]

La Liga or formally the Primera División is the top of the Spanish professional football league system. It was founded in 1929 and has hosted 59 teams. For a long period of time the championship has been dominated by Real Madrid and F.C. Barcelona with 32 and 22 titles respectively. However, in the past two decades other teams have earned the title as well.[39]

3.1 Data Retrieval and Format

The dataset has been created from football and odds statistics extracted from two websites providing freely available football data. These are:

- `www.football-data.co.uk`
- `www.betexplorer.com`

The first website offers free download of data files (.csv format) covering odds data and match data from the two leagues from seasons 00/01-12/13. Odds from 13 different bookmakers have been collected on the website, although at most 9 are present in single seasons. The relevant data in the files has been extracted and converted into .mat files in MATLAB.

The second website does not offer downloadable data files with odds and match results. Instead it has a comprehensive free database of odds stored in tables on numerous sub pages. The website offers odds from 47 different bookmakers from seasons 08/09-12/13. Not all bookmakers are present in all matches. To easily retrieve and update all data, an API has been implemented in MATLAB to download and store the data as .mat files.

For each season in each of the two leagues, the two data sources have been merged to create single .mat files. Whenever odds are present in both data sources, the odds from the latter and also largest data source is used. The dataset has been cleared of prospective outliers by discarding any odds which deviate by more than 75% from the mean odds value. Let o_{ij}^k, $i = 1, \ldots, M_t$,$j = 1, 2, 3$, $k = 1, \ldots N_t$ denote the odds value of the kth bookmaker in the jth selection in match i, where M_t and N_t is the total number of available

matches and bookmakers respectively. Further let $\bar{o}_{ij} = \frac{1}{N_t}\sum_{k=1}^{N_t} o_{ij}^k$ denote the mean odds value among the N_t bookmakers. The required criteria to retain $[o_{i1}^k, o_{i2}^k, o_{i3}^k]$ for bookmaker k can then be formalized as

$$\frac{|\bar{o}_{ij} - o_{ij}^k|}{\bar{o}_{ij}} \leq 0.75, \quad j = 1, 2, 3 \tag{3.1}$$

In the PL dataset 22 books and 1 book have been rejected in seasons 08/09 and 12/13 respectively. In the LL dataset 23, 9 and 4 books have been rejected in seasons 08/09, 11/12 and 12/13 respectively.

The basic content of the data set, one for each league and for each season, can then be summarized as:

- Full time result: Home win (1), draw (2) away win (3), stored in an $M_t \times 1$ array.
- Full time result odds from different bookmakers: Home odds o_1, draw odds o_2 and away odds o_3, stored in an $M_t \times (3 \cdot N_t)$ array.
- Opposing teams in each match, stored in an $M_t \times 1$ cell structure.

Missing or removed odds values are replaced by empty elements. The times at which the odds have been registered on the websites are unclear. On `football-data.co.uk` there is no information available. `betexplorer.com` offers the possibility of tracking the odds movement about 24 hours before each match at the current season 12/13. A random check of different bookmakers and matches shows that most odds have been registered about 1-4 hours before the match kick-off.

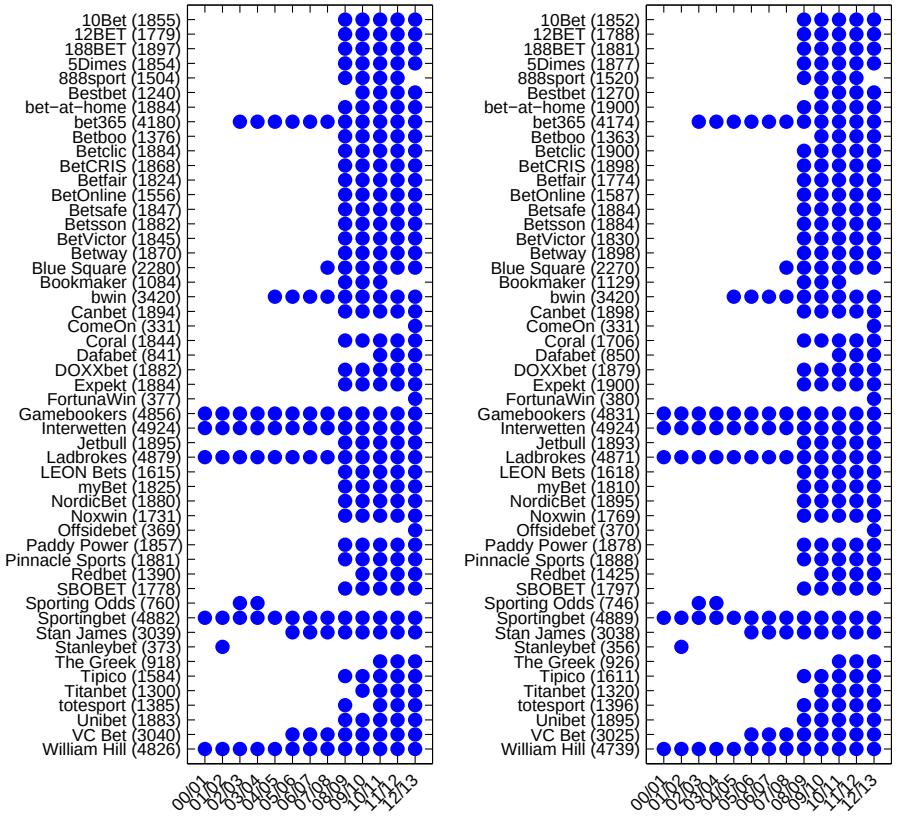
3.2 Data Content

Figures 3.1a and 3.1b depict the presence of each bookmaker across seasons 00/01-12/13 in PL and LL respectively. Seasons 00/01-07/08 only contain data from `football-data.co.uk`, which only offers odds from few bookmakers compared to `betexplorer.com`. This explains the significant change in the number of registered bookmakers in seasons 07/08-08/09. It should also be emphasized that many of the bookmakers were not founded before the later seasons including 188Bet (founded in 2006), Betsafe (2006), FortunaWin (2009), Jetbull (2007), Leon Bets (2007), Noxwin (2007) and Titan Bet (2009). The most

covered bookmakers which are present in all seasons in both leagues are Gamebookers (1998), Interwetten (1990), Ladbrokes (1886), Sporting bet (1998) and William Hill (1934) which are also among the oldest bookmakers. [40] Evidently these bookmakers are not present in all matches as they are exceeded by the theoretical maximum number of odds per bookmaker, $380 \cdot 13 = 4940$. A closer examination of the datasets reveal that many bookmakers have missing matches in seasons where they are present. Accordingly, the number of matches covered by all bookmakers is very limited, and in order to obtain a good coverage with a given number of bookmakers $N \leq N_t$ it would seem reasonable to collect odds from the N most recorded bookmakers. For the remainder of the thesis only $N = 5$, $N = 9$ and $N = 37$ will be considered, as they provide a good odds coverage relative to N, see Table 3.1. In case $N = 5$ these are: Gamebookers, Interwetten, Ladbrokes, Sportingbet, and William Hill. In case $N = 9$ these are: bet365, bwin, Gamebookers, Interwetten, Ladbrokes, Sportingbet, Stan James, VC Bet, and William Hill. In case $N = 37$ this corresponds to all the 37 bookmakers present in seasons 08/09-12/13 in both leagues.

		No missing odds		
League	*Missing odds*	$N = 5$	$N = 9$	$N = 37$
Premier League	4940	4647	2929	1160
La Liga	4938	4567	2913	1039

Table 3.1: Number of matches with and without removal of matches with missing odds, seasons 00/01-12/13. N refers to the number bookmakers with most individually covered matches.



(a) Premier League **(b)** La Liga

Figure 3.1: Presence of bookmakers according to season. The parentheses state the total number of odds for each bookmaker. Note that the number of present bookmakers is significantly larger in the latest five seasons, as these seasons are covered by both data sources.

3.3 Odds Analysis

3.3.1 Outcome frequencies

Table 3.2 shows the distribution among the outcomes home (1), draw (2) and away (3), and the distribution among the bookmakers' favourite selections. The odds have been collected over all 13 seasons available. The favourite selections

have been determined as the the selection with smallest mean odds value. Let o_{ij}^k, $i = 1, \ldots, M$, $j = 1, 2, 3$, $k = 1, \ldots N$ denote the odds value of the kth bookmaker in the jth selection in match i, where M and N are the numbers of matches and bookmakers respectively. The favourite selection $\tilde{T}_i = 1, 2, 3$ in match i can then be formalized as

$$\tilde{T}_i = \arg\min_j \overline{o_{ij}}, \quad i = 1, \ldots, M \tag{3.2}$$

where

$$\overline{o_{ij}} = \frac{1}{N} \sum_{k=1}^{N} o_{ij}^k \tag{3.3}$$

In Table 3.2 all $N_t = 51$ bookmakers have been used and consequently $M_{PL} = 4940$ and $M_{LL} = 4938$. Missing odds are omitted when evaluating the mean odds for each match. Evidently the home teams are biased to win, as almost 50% of all matches are won by the home team. This strongly emphasizes the importance of the home team advantage as a factor of the overall performance of each team. In both leagues draws and away wins are overall equally likely, as each of these outcomes account for approximately 25% of all outcomes.

	Wins [%]			*Favourites* [%]		
League	H	D	A	H	D	A
Premier League	46.7	26.1	27.2	74.0	0.00	26.1
La Liga	48.5	24.8	26.8	79.1	0.00	20.9

Table 3.2: Distribution of match results (Wins) and selections favoured by the bookmakers (Favourites), seasons 00/01-12/13. The favourite in each match is determined as the selection with smallest mean odds.

The bookmakers consistently never favour the draw selection even though almost 25% of all matches are draws. The away selection is generally also less favoured than the actual frequency of away wins. Consequently, the home teams are heavily favoured compared to the actual frequency of home wins, which indicates a strong home team bias in the odds. In summary two levels of bias are observed: A competitive home ground bias and a home team favouring bias in the odds, whose cumulative effects lead to a distinct bookmaker favouring of home selections.

3.3.2 Odds distribution

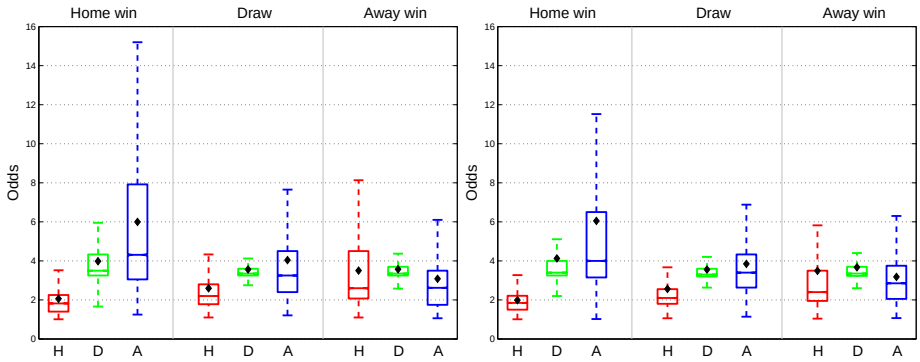
The books from each of the two leagues have been divided into three subsets containing books from home wins, draws and away wins. Specifically the data including all bookmakers and seasons has been divided into three subsets

$$\boldsymbol{O}_{\text{home}} = \{\left[o_{i1}^k, o_{i2}^k, o_{i3}^k\right]^T \mid T_i = 1\}, \tag{3.4}$$

$$\boldsymbol{O}_{\text{draw}} = \{\left[o_{i1}^k, o_{i2}^k, o_{i3}^k\right]^T \mid T_i = 2\}, \tag{3.5}$$

$$\boldsymbol{O}_{\text{away}} = \{\left[o_{i1}^k, o_{i2}^k, o_{i3}^k\right]^T \mid T_i = 3\}, \quad i = 1, \ldots, M_t, \quad k = 1, \ldots, N_t \tag{3.6}$$

Figure 3.2 depicts the odds distribution on each selection in $\boldsymbol{O}_{\text{home}}$, $\boldsymbol{O}_{\text{draw}}$ and $\boldsymbol{O}_{\text{away}}$ in both leagues. Each selection contains significant outliers regardless of the outcome which essentially capture the statistically uneven matches with extreme single odds values. For illustration purposes the outliers have been removed. Supplementary statistics are summarized in Table 3.3.



(a) Premier League. **(b)** La Liga.

Figure 3.2: Odds distribution by selection, seasons 00/01-12/13.

In general all distributions are strongly right-skewed, as there are numerous extreme odds values identified as outliers. In home win matches the distribution of the home odds indicates that the home team is generally favoured and that this is a persistent perception, as the variance is very low. Contrary there is no clear favouring of the away team in away win matches. In fact the home and away teams are considered to be more or less equal, emphasizing the immense effect of the home ground advantage. The odds on away selections have a generally higher variance than the odds on the other two selections, suggesting that the *probabilistic span* implied by the bookmakers' odds is significantly larger on the away team. Specifically a very weak away team will have a far lower

	Home win					
	PL			*LL*		
	mean	*median*	*var.*	*mean*	*median*	*var.*
Home	1.99	1.85	0.970	2.06	1.82	1.12
Draw	4.13	3.40	3.25	3.98	3.50	1.29
Away	6.05	4.00	29.0	6.00	4.31	18.9
	Draw					
	PL			*LL*		
	mean	*median*	*var.*	*mean*	*median*	*var.*
Home	2.56	2.10	3.20	2.60	2.20	1.95
Draw	3.55	3.30	0.548	3.56	3.35	0.373
Away	3.84	3.40	4.80	4.04	3.25	6.90
	Away win					
	PL			*LL*		
	mean	*median*	*var.*	*mean*	*median*	*var.*
Home	3.49	2.40	8.09	3.50	2.60	4.62
Draw	3.67	3.35	0.738	3.57	3.35	0.347
Away	3.18	2.85	3.78	3.08	2.62	4.16

Table 3.3: Basic statistics on the Premier League (PL) and La Liga (LL) datasets, seasons 00/01-12/13 on home, draw and away selections conditioned on the type of outcome.

chance of success than a very weak home team, according to the probabilities implied by the odds.

The draw and away odds are predominantly large on home wins compared to draw and away wins. In fact the sum of odds averaged over all bookmakers and all matches in each of the sets $\boldsymbol{O}_{\text{home}}$, $\boldsymbol{O}_{\text{draw}}$ and $\boldsymbol{O}_{\text{away}}$ are 12.0, 10.2, and 10.2 for the PL dataset and 12.1, 9.95 and 10.3 for the LL dataset, respectively. This indicates that odds are generally higher on home wins and could be perceived as more attractive. However, the average overrounds are very similar valued at 0.07 in the PL dataset and 0.08 in the LL dataset regardless of the outcome. Evidently the odds on home wins are simply distributed in a convenient way that allows high odds on single selections while maintaining an indifferent overround.

3.3.3 Temporal development of overround

Figure 3.3 depicts the temporal development of the overround across all available seasons with $N = 5$ bookmakers. A set of N overrounds, one for each bookmaker, has been extracted from each match. Evidently the average overround is monotonically decreasing as a function of the seasons, and has nearly halved during the past 13 seasons. The clear tendency is presumably a result of increased market competition, effectively augmenting the odds values. Ref. [20] demonstrated that the overround was remarkably constant at around 0.115, when evaluating on odds from an unspecified set of bookmakers on on 3382 matches from two English leagues (1993-1994). Thus the observed overrounds from the current dataset at hand, indicates slightly higher overrounds on early seasons than demonstrated by [20].

Comparing the two leagues reveals that the overround is slightly lower on the PL dataset in all seasons, cf. Table 3.4, which suggests that the odds are generally more attractive in the Premier league compared to the La Liga.

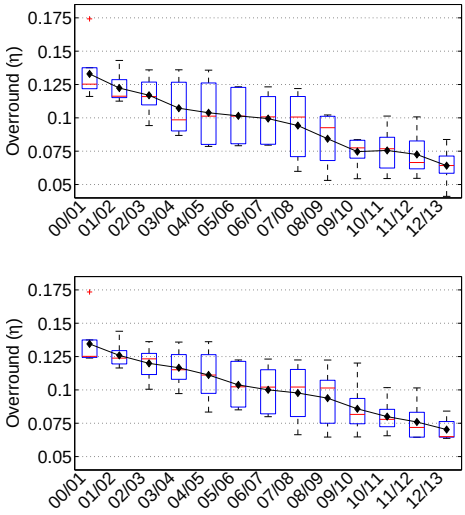


Figure 3.3: Temporal development of overround. Top: Premier League (PL), bottom: La Liga (LL).

	Overround	
Season	*PL*	*LL*
00/01	0.1329	0.1345
01/02	0.1223	0.1259
02/03	0.1170	0.1199
03/04	0.1073	0.1166
04/05	0.1038	0.1112
05/06	0.1015	0.1037
06/07	0.0995	0.1001
07/08	0.0942	0.0977
08/09	0.0843	0.0938
09/10	0.0747	0.0858
10/11	0.0755	0.0800
11/12	0.0725	0.0759
12/13	0.0641	0.0703

Table 3.4: Mean overround in PL and LL per season.

3.3.4 Spatial odds distribution by class

One of the central challenges is to discriminate between the characteristics of the odds conditioned on the three different outcomes. In order to effectively extract and visualize the data information a principal component analysis (PCA) has been performed on both datasets. Each match i is represented by a $3{\cdot}N{\times}1$ array containing $[o_{i1}, o_{i2}, o_{i3}]^T$, $j = 1, \ldots, N$. With $N = 5$ a total of $M_{PL} = 4647$ and $M_{LL} = 4567$ matches (no missing values) have been recorded on the PL and LL sets respectively, c.f. Table 3.1, and the data projections onto the first 3 principal components (PCs) are depicted in figures 3.5 and 3.4.

The explained cumulative variance ([%]) with one, two and three PCs, are 0.6726, 0.9652, and 0.9728 for the PL dataset and 0.6272, 0.9697, and 0.9767 for the LL dataset, respectively. The shape of the data combined with the amount of explained variance indicate that the features stem from a one-dimensional manifold in a two-dimensional space. Hence the inherent complexity of the odds has two degrees of freedom. This observation is reasonable considering that 1) the bookmakers' odds are presumably very similar, which effectively reduces the degrees of freedom to 3; 2) the overround remains reasonably constant, which further reduces the degrees of freedom to 2. Therefore the two first PCs primarily address the *general* odds structure, whereas the remaining variance, primarily explained by the third PC, explains the *distinction* between the bookmakers.

Quite interestingly the samples are arranged in a U-shape, where the two end segments are dominated by the home and away win classes respectively. The odds from the draw class are located all around in the data cloud. Most samples are concentrated in the 'bend' of the data cloud, corresponding to the matches where there is no clear team favourite, which also comprises the majority of matches. In both end segments the remaining two classes are sporadically present, corresponding to matches in which the bookmakers' anticipations strongly conflict with the outcome. A careful inspection of the plots reveals considerably more of these outliers are present in the home class end segment in the PL dataset, indicating that more matches results in draws despite significant home team favourings.

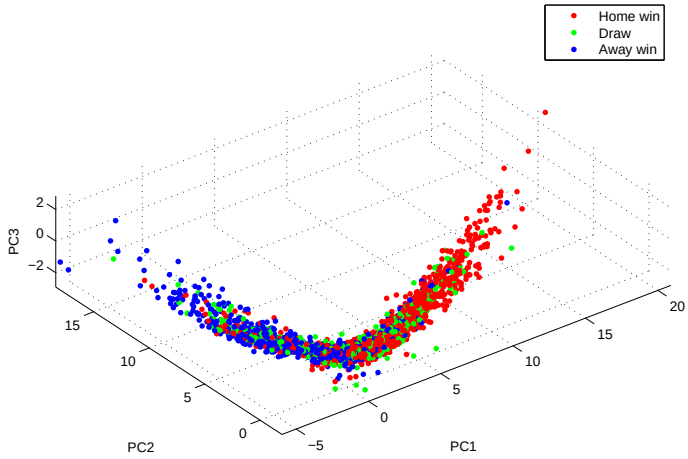


Figure 3.4: PCA on Premier League dataset, seasons 00/01-12/13. Each sample (match) consists of the odds from the 5 most covered bookmakers.

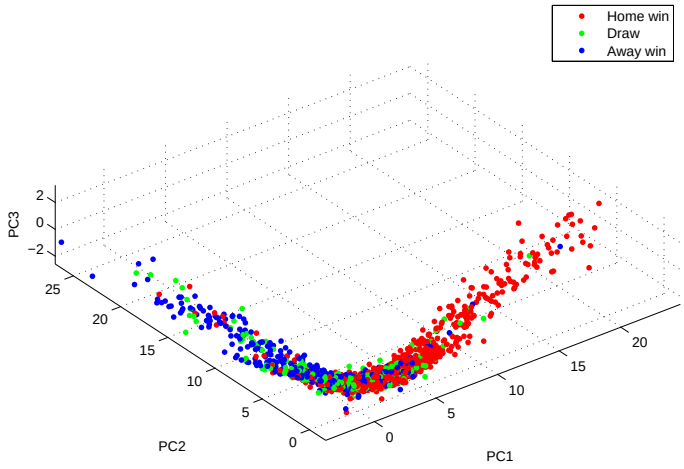


Figure 3.5: PCA on La Liga dataset, seasons 00/01-12/13. Each sample (match) consists of the odds from the 5 most covered bookmakers.

3.4 Bookmaker Comparison

3.4.1 Overround comparison

Figure 3.6 depicts the distribution of the overround η of the 37 most recorded the LL dataset in seasons 08/09-12/13. Similar distributions are obtained using the PL dataset. η has been collected using Eq. (2.5) on all odds, i.e. $\left[o_{i1}^k, o_{i2}^k, o_{i3}^k\right]^T$, $i = 1, \ldots, M$, $j = 1, \ldots, 37$. Missing odds are omitted. So each bookmaker is covered differently in the matches, cf. figure 3.1b. Quite remarkably η varies a lot amongst many bookmakers and the mean overround is different. Pinnacle Sports has the lowest mean overround of $\overline{\eta} = 0.0218$ closely followed by 5Dimes (0.0245), 188BET (0.0360) and Betfair (0.0380). BetCRIS holds the highest mean overround of $\overline{\eta} = 0.1022$ followed by Interwetten (0.0944) and Sportingbet (0.0910). While one may question how the latter are competitive with the former it should be emphasized that the overround is a combined measure over all three selections and does not necessarily reflect the attractiveness on specific selections in specific matches. The variety of overround distributions indicates that the bookmakers use very different strategies to determine their odds, and suggests that each bookmaker encompasses different latent features of each match.

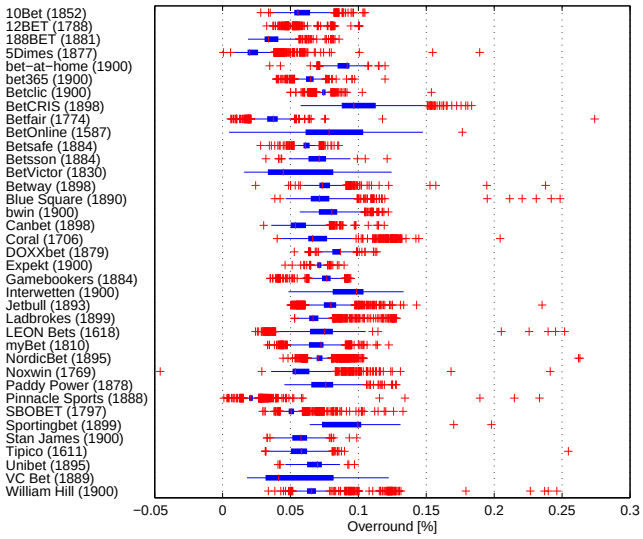


Figure 3.6: Distribution of overround of the 37 most covered bookmakers in the La Liga dataset, seasons 08/09-12/13. The brackets hold the sample count.

3.4.2 Cross-comparison of bookmakers

The $N = 9$ most recorded bookmakers are compared against each other to identify possible atypical bookmakers and the main sources of the variation.

First the correlation matrix $\boldsymbol{R} \in \mathbb{R}^{9\times 9}$ has been considered. The correlations are determined as the mean correlations across each outcome class. Specifically each element $\boldsymbol{R}_{mn}$, is found as

$$\boldsymbol{R}_{mn} = \frac{1}{3}\left(\frac{\boldsymbol{\sigma}_{mn,h}}{\boldsymbol{\sigma}_{m,h}\boldsymbol{\sigma}_{n,h}} + \frac{\boldsymbol{\sigma}_{mn,d}}{\boldsymbol{\sigma}_{m,d}\boldsymbol{\sigma}_{n,d}} + \frac{\boldsymbol{\sigma}_{mn,a}}{\boldsymbol{\sigma}_{m,a}\boldsymbol{\sigma}_{n,a}}\right), \quad m,n = 1,\ldots,9 \tag{3.7}$$

where

$$\boldsymbol{\sigma}_{mn,h} = \mathrm{Cov}(\boldsymbol{o}^m_{1,h}, \boldsymbol{o}^n_{1,h}), \qquad \boldsymbol{\sigma}_{n,h} = \sqrt{\mathrm{Var}(\boldsymbol{o}^n_{1,h})} \tag{3.8}$$

$$\boldsymbol{\sigma}_{mn,d} = \mathrm{Cov}(\boldsymbol{o}^m_{1,d}, \boldsymbol{o}^n_{1,d}), \qquad \boldsymbol{\sigma}_{n,d} = \sqrt{\mathrm{Var}(\boldsymbol{o}^n_{1,d})} \tag{3.9}$$

$$\boldsymbol{\sigma}_{mn,a} = \mathrm{Cov}(\boldsymbol{o}^m_{1,a}, \boldsymbol{o}^n_{1,a}), \qquad \boldsymbol{\sigma}_{n,a} = \sqrt{\mathrm{Var}(\boldsymbol{o}^n_{1,a})} \tag{3.10}$$

where $\boldsymbol{o}^n_{1,h}, \boldsymbol{o}^n_{1,d}, \boldsymbol{o}^n_{1,a} \in \mathbb{R}^M$ contain all home, draw and away odds, respectively, for the nth bookmaker in temporal order. Figure 3.7 depicts the correlation using the LL dataset, seasons 05/06-12/13. All bookmakers are highly correlated, but some bookmakers stand out slightly. Especially VC Bet and Interwetten differ from the remaining 8 bookmakers, and particularly they differ from each other. Stan James and bet365 have the highest correlation with $\rho = 0.984$.

The contributions to the differences between bookmakers' odds have been separated on the 3 selections. Figure 3.8 illustrates the mean absolute distance e between the 9 bookmakers on each odds selection, i.e. $e_{mn,j} = \frac{1}{M}\sum_{i=1}^{M}|o^m_{ij} - o^n_{ij}|$, where $j = 1, 2, 3$ indicates the selection on the LL dataset, seasons 05/06-12/13. Evidently the largest error contribution comes from the away odds with a mean residual of $\bar{e}_3 = 0.4033$. The bookmakers are very indifferent on the home and draw odds, as the mean residuals are $\bar{e}_1 = 0.1285$ and $\bar{e}_2 = 0.1729$ respectively. In accordance with figure 3.7 VC Bet and Interwetten differs much from the remaining bookmakers an in particular from each other. Also Stan James differs from the rest, except from bet365.

In order to obtain a global distance measure between each of the 9 bookmakers a PCA has been performed on 9 samples on the LL dataset, seasons 05/06-12/13, containing odds from all matches where all 9 bookmakers are present. The odds $\left[o^k_{i1}, o^k_{i2}, o^k_{i3}\right]^T$, $i = 1,\ldots,M$ for $k = 1,\ldots,9$ have been stacked into single vectors of size $3 \cdot M = 3 \cdot 2825 = 8739$. Figure 3.9 displays the first three principal components, explaining 52.6% of the data variation.

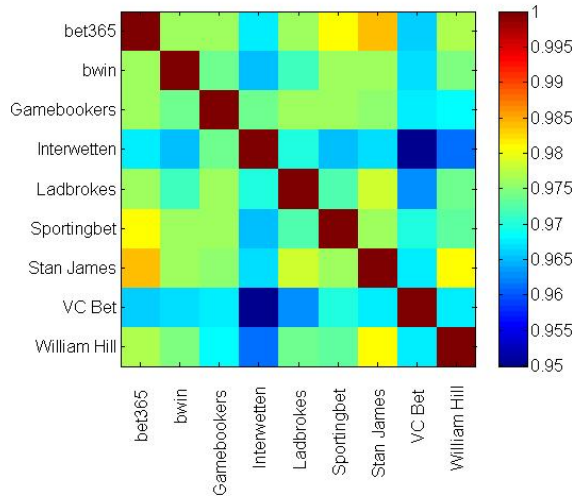


Figure 3.7: Correlation between the 9 most covered bookmakers from the La Liga dataset, seasons 05/06-12/13.

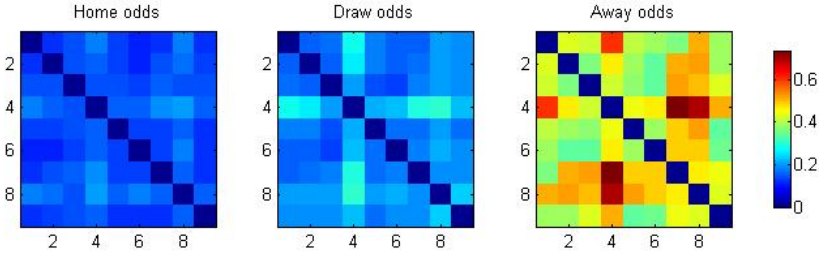


Figure 3.8: Mean residuals between the 9 most covered bookmakers from the La Liga dataset, seasons 05/06-12/13. The indices on the axes refer to the bookmakers in figure 3.7.

Since the PCA decomposes the covariance of the data, the plot indicates to which extend the samples vary together. Evidently the bookmakers have very low covariances due to the poor amount of variance explained by the first PCs. Comparing the low covariance with the high correlation observed figure 3.7, suggests that the odds setting of each bookmaker strongly affects the odds setting of the rest of the bookmakers. Although the bookmakers generally agree on the magnitude of the odds the individual odds strategies are highly separable, expressed by the low covariance. This is an important insight as there now

is strong evidence that bookmakers constitute a committee of partly independent experts which can be utilized to obtain robust estimates of the posterior probabilities on each selection, given the odds from the various bookmakers.

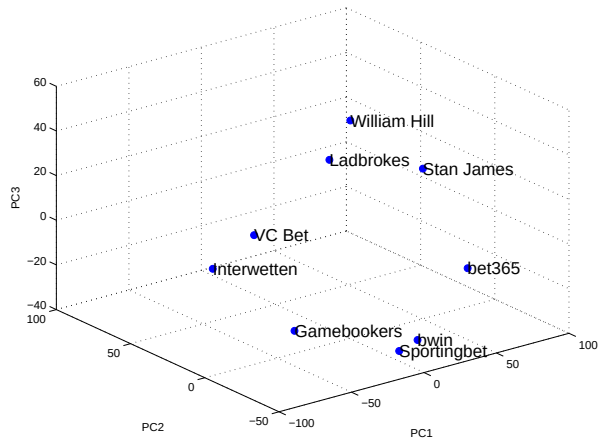


Figure 3.9: PCA on the 9 most covered bookmakers from La Liga dataset, seasons 05/06-12/13. Each bookmaker is represented by all its odds, corresponding to a 8739 dimensional data point.

3.5 Team Bias

An natural question that arises is whether the probability implied by the odds reflects the true probability of a given teams success or whether the effects of odds balancing, particularly the favourite/long-shot bias, influence the implied probability on the individual teams. This question will be answered by estimating the empirical probability of a given teams success with the probability implied by the odds. To clarify the procedure the following formalism is introduced. Let $\tilde{p}_{ij}$ denote the implied mean probability among N bookmakers in match i on selection j defined as

$$\tilde{p}_{ij} = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{o_{ij}^k (1 + \eta_i^k)} \tag{3.11}$$

where η_i^k is the overround of bookmaker $k = 1, \ldots, N$ in match $i = 1, \ldots, M$, defined as (cf. Eq. (2.5))

$$\eta_i^k = \left(\sum_{j=1}^{3} \frac{1}{o_{ij}^k} \right) - 1 \tag{3.12}$$

Further let δ_{ij} define a binary function indicating whether the jth selection in match i came true or not,

$$\delta_{ij}(x) = \begin{cases} 1, & T_i = j \\ 0, & \text{otherwise} \end{cases} \tag{3.13}$$

where $T_i = 1, 2, 3$ states the outcome of the ith match. Accordingly, δ_{ij} can be regarded as the empirical probability for the jth selection in the ith match, with a probability of either 0 or 1. Now let $S_{kl}^{\tilde{p}}$ denote the set containing all probabilities $\tilde{p}_{ij} \in I_k$, and S_{kl}^{δ} corresponding binary value δ_{ij} for team $l = 1, \ldots, L$ among all L recorded teams, with $j = 1, 3$, $i = 1, \ldots, M$, $k = 1, \ldots, 10$ and where

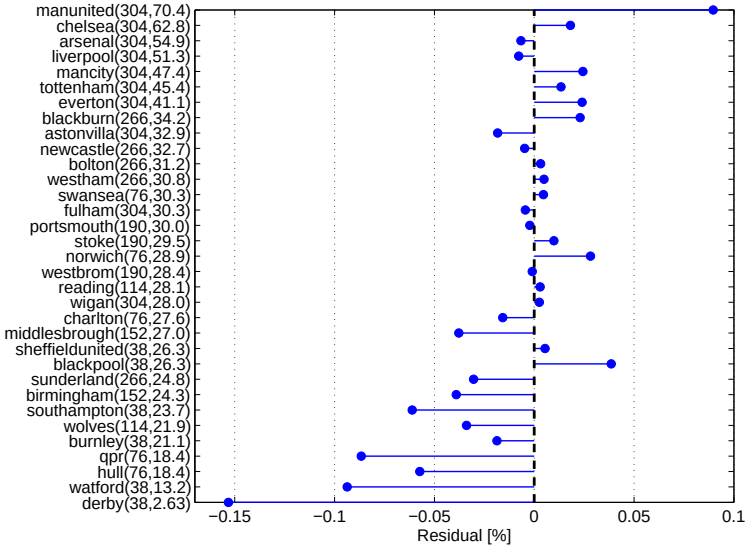
$$I_1 = [0, 0.1[,\ I_2 = [0.1, 0, 2[,\ \ldots,\ I_{10} = [0.9, 1] \tag{3.14}$$

denote subintervals of $[0, 1]$. A reasonable measure of deviation between the empirical (true) probabilities and implied probabilities can then be found as the weighted mean value of the difference between mean success rate, $\mathrm{E}[S_{kl}^{\delta}]$, and mean implied probability, $\mathrm{E}[S_{kl}^{\tilde{p}}]$, across the k subintervals. To put it explicitly,

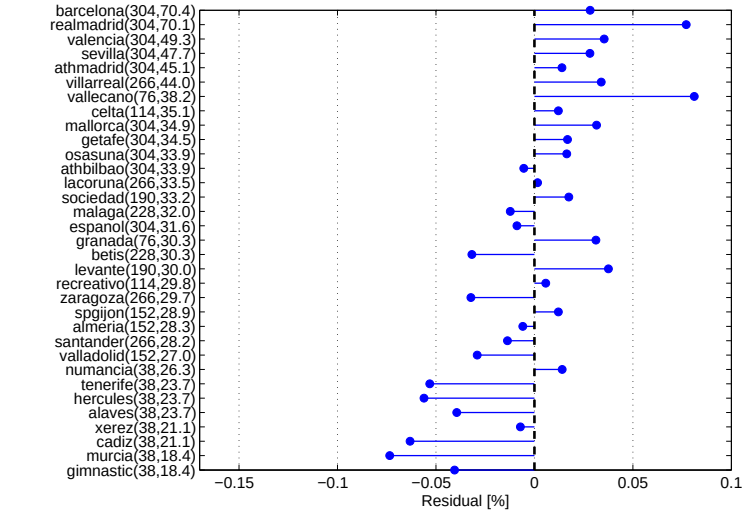
$$r_l = \frac{\sum_{k=1}^{10} N_{kl}(\mathrm{E}[S_{kl}^{\delta}] - \mathrm{E}[S_{kl}^{\tilde{p}}])}{\sum_{k=1}^{10} N_{kl}}, \quad l = 1, \ldots, L \tag{3.15}$$

where N_{kl} it the number of elements in the set S_{kl}^{δ}.

Figures 3.10a and 3.10b depict the residuals on the PL and LL datasets respectively, seasons 05/06-12/13 with $N = 9$ including all participating teams, ordered according to success rate. If $r_l > 0$ it will indicate that team l is generally underestimated and vice versa if $r_l < 0$. Neglecting the smallest (noisy) residuals there is a tendency towards underestimating the strongest teams while overestimating the weakest teams. In particular the superior performances of Manchester United and Real Madrid are heavily understated, while the performances of e.g. QPR and Hercules are heavily overestimated. Although some teams are even stronger overestimated these teams are only weakly represented due to few active seasons, and should therefore be processed with caution. These observations demonstrate that the favourite/long-shot bias cf. section 2.3.3 indeed is prevalent in both leagues and suggest the presence of significant team loyalty on small, unfavoured teams.



(a) Premier League.



(b) La Liga.

Figure 3.10: Estimated differences between empirical probabilities and probabilities implied by the odds, seasons 05/06-12/13. The brackets hold (#matches, success rate).

Chapter 4

Model Definition

The proposed model is based on a neural network (NN) for multi-class classification combined with a decision framework based on determining the expected returns on different combinations of engaged selections per match. Section 4.1 provides an overview of the basic components and procedures involved in application of NNs, section 4.2 a brief description of the applied MATLAB toolbox, and section 4.3 gives an explicit formulation of the input features and the decision framework.

4.1 Artificial Neural Networks for Multi-class Classification

Artificial Neural Networks (NN) constitute a class of highly flexible models for regression problems and classifications problem. The term 'neural network' originates from the attempts to mimic the behaviour in biological systems, particularly the cognitive structure of the brain and are widely used in pattern recognition.

A NN admits itself to a feed-forward architecture, where the input signal propagates though a set of (non-linear) functional transformations which are linked

through a directed acyclic graph to ensure deterministic outputs of the inputs. NNs consist of layers of neurons; an input layer, one or more hidden layers, and an output layer. For the sake of relevance only NNs with one hidden layer are considered with full links between the input and hidden layers, and the hidden and output layers. Figure 4.1 depicts a NN with four input units, five hidden units in a single layer and one output unit.

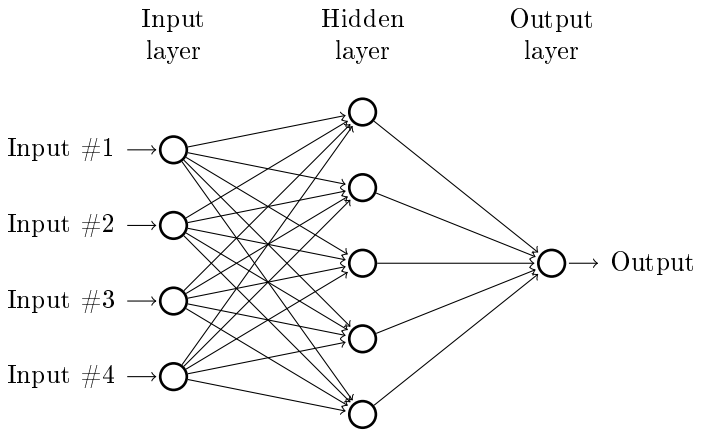


Figure 4.1: An example of a neural network with five hidden units in one hidden layer, four input units, and one output unit. For simplicity, the bias parameters have not been included.

The input variables x_i, $i = 1, \ldots, N_x$ are passed on to the hidden layer by N_z linear combinations, which are transformed using a non-linear activation function $h(\cdot)$,

$$z_k = h(a_k^h) = h\left(\sum_{j=0}^{N_x} w_{kj}^h x_j\right), \quad k = 1, \ldots, N_z \tag{4.1}$$

where z_k it the activation energy in the kth hidden unit, N_z is the number of hidden units, a_k^h refers to the activation of the kth hidden unit, and w_{kj}^h to the weight of the link between the jth input and the kth hidden unit. For convenience the indexing begins at zero to encompass the bias/threshold parameters w_{0j}^h , as x_0 is fixed at $x_0 = 1$. For multi-class classification the posterior probabilities (output units) y_k for a given input sample $\boldsymbol{x}_k$ are given by the softmax function of linear combinations of the hidden unit variables,

$$p(\mathcal{C}_k|\boldsymbol{x}, \boldsymbol{w}) = y_k = \frac{\exp a_k^o}{\sum_{j=0}^{Nz} \exp a_j^o}, \quad k = 1, \ldots, N_y \tag{4.2}$$

where N_y is the number of output units (or classes), $\boldsymbol{w}$ is the set of all weights, and the output activations a_k^o are given by

$$a_k^o = \sum_{j=0}^{N_z} w_{kj}^o z_j, \quad k = 1, \ldots, N_y \tag{4.3}$$

where w_{kj}^o is the weight of the link between the jth hidden and the kth output unit. It is noted that $\sum_{k=1}^{N_y} y_k = 1$ and $y_k \in [0,1]$, $\forall k \in 0, \ldots, N_y$ whereby y_k indeed have the characteristics of class posteriors. As with Eq. (4.1) the indexing starts at zero to encompass the bias parameters w_{0j}^o, as $z_0 = 1$.

The task is now to optimize the weights $\boldsymbol{w}$, by maximizing the likelihood. Introduce a binary target variables $t_{nk} \in \{0,1\}$ with a 1-of-C coding scheme to indicate the class of input sample $\boldsymbol{x}_n \in \mathbb{R}^{N_x}$ among C mutually exclusive classes. Assuming independence between class labels and input samples, the likelihood can then be formulated as

$$p(T|\boldsymbol{w}, \boldsymbol{x}) = \prod_{n=1}^{N_{\text{samples}}} \prod_{k=1}^{C} p(\mathcal{C}_k|\boldsymbol{w}, x_n)^{t_{nk}} = \prod_{n=1}^{N_{\text{samples}}} \prod_{k=1}^{C} y_k^{t_{nk}} \tag{4.4}$$

where t_{nk} is a 1-of-K coding scheme and T is a $N_{\text{samples}} \times C$ matrix with elements t_{nk}, and N_{samples} is the number of input samples. Eq. (4.4) is essentially a product of the 'active' (independent) posteriors which should be maximized. For convenience the likelihood is transformed leading to an equivalent minimization problem, expressed by the cross-entropy error function for multi-class

$$E(\boldsymbol{w}) = -\ln p(T|\boldsymbol{w}, \boldsymbol{x}) = -\sum_{n=1}^{N_{\text{samples}}} \sum_{k=1}^{C} t_{nk} \ln y_k \tag{4.5}$$

Since the error function is smooth, its minimum will occur in the weight space where the gradient is zero, i.e. $\nabla E(\boldsymbol{w}) = 0$. Evidently there is no chance of finding an analytical solution to the problem, so it should be solved numerically by non-linear numerical optimization techniques which predominantly strive to optimize the parameters $\boldsymbol{w}$ in the weight space though iterative updating procedures. The simplest approach is gradient descend, where the gradient information is utilized to choose the weight update opposite of the gradient direction,

$$\boldsymbol{w}^{(\text{new})} = \boldsymbol{w}^{(\text{old})} - \lambda \nabla E(\boldsymbol{w}^{(\text{old})}) \tag{4.6}$$

where λ denotes the learning rate or step size. [22, pp. 226-240]

An efficient technique for evaluating $\nabla E(\boldsymbol{w})$ can be achieved by using a local information passing scheme commonly known as back-propagation. By dividing

the error function into a sum of terms, one for each input sample, i.e.

$$E(\boldsymbol{w}) = \sum_{n=1}^{N_{\text{sample}}} E_n(\boldsymbol{w}) \tag{4.7}$$

it can be shown that the partial derivatives with respect to $\boldsymbol{w}$ can be found as

$$\frac{\delta E_n}{\delta w_{ji}^o} = \delta_j^o z_i = (y_j - t_j) z_i \tag{4.8}$$

$$\frac{\delta E_n}{\delta w_{ji}^h} = \delta_j^h z_i = \left(h'(a_j^h) \sum_{k=0}^{N_y} w_{kj}^o \delta_k^o \right) x_i \tag{4.9}$$

where the δs, often referred to as errors, are a useful notation to promote the transparency and efficiency of the calculations, as δ_j^o is reused in both Eqs. (4.8) and (4.9). See Appendix C.1 for a derivation of the partial derivatives. Eq. (4.9) reveals that the partial derivatives with respect to w_{ji}^h are found as a backwards propagation of the errors δ_k^o from the output layer, contrary to the forward propagation found in Eqs. (4.1) and (4.2) [22, pp. 240-244]. Intuitively this makes sense, as the magnitude of $\frac{\delta E_n}{\delta w_{ji}^h}$ is determined by the magnitude of the each error $\delta_j^o = y_j - t_j$ from the output layer, weighted by the strength of the individual error signals, i.e. weighted by w_{kj}^o.

While the gradient descend method only applies first order derivatives, a variety of potentially more efficient second order algorithms are available as well, such as Newton, Levenberg-Marquardt, Gauss-Newton and pseudo-Gauss-Newton. These can be developed from a second order Taylor expansion of the cost function $E(\boldsymbol{w})$ around a point $\hat{\boldsymbol{w}}$ in the weight space,

$$E(\boldsymbol{w}) = E(\hat{\boldsymbol{w}}) + (\boldsymbol{w} - \hat{\boldsymbol{w}})^T \nabla E(\hat{\boldsymbol{w}}) + \frac{1}{2} (\boldsymbol{w} - \hat{\boldsymbol{w}})^T \boldsymbol{H}_{\hat{\boldsymbol{w}}} (\boldsymbol{w} - \hat{\boldsymbol{w}}) \tag{4.10}$$

where $\boldsymbol{H}_{\hat{\boldsymbol{w}}}$ denotes the Hessian at $\hat{\boldsymbol{w}}$. Similarly to the gradient $\nabla E(\boldsymbol{w})$, the Hessian can be efficiently evaluated by means of the back propagation procedure The gradient is vanishing at a (local) minimum $\boldsymbol{w}_0$, i.e.

$$\nabla E(\boldsymbol{w}_0) = \nabla E(\hat{\boldsymbol{w}}) + \boldsymbol{H}_{\hat{\boldsymbol{w}}} (\boldsymbol{w}_0 - \hat{\boldsymbol{w}}) = \boldsymbol{0} \tag{4.11}$$

which implies that

$$\boldsymbol{w}_0 = \boldsymbol{w} - \boldsymbol{H}_{\hat{\boldsymbol{w}}}^{-1} \nabla E(\hat{\boldsymbol{w}}) \tag{4.12}$$

Accordingly, the inverse of the Hessian is required, which may cause numerical difficulties. [41, pp. 4-5] For this reason there have been interests in using the diagonal Hessian instead, as its inverse is easily evaluated. However, the Hessian is often strongly non-diagonal and so the diagonal Hessian should be used with care [22, p. 250].

4.2 DTU Neural Classification Toolbox

The applied toolbox, `nc_toolbox`, is an older version of the current toolbox `nc_multiclass` provided by DTU Compute, which can be applied freely in research and other non-profit applications.[1] The older version has been applied as the computation time is significantly lower than with the newer toolboxes, while the probabilistic accuracy of the toolboxes are indifferent to the classification problem at hand.

`nc_toolbox` uses the hyperbolic tangent function as activation function, cf. Eq. (4.1) and applies gradient descent followed by pseudo-Gauss-Newton using a diagonal Hessian approximation. The program has been set to perform 10 gradient descend iterations and 50 pseudo-Gauss-Newton iterations unless a gradient norm stopping criteria of $\lambda = 10^{-4}$ has been reached. The 'optimal' step size λ in each iteration is determined by a simple line search in the weight space with iterative bisection. Although the implementation offers adjustment of regularization parameters, controlling the learning rate on the weights $\boldsymbol{w}$, these have been set to the default values. A further analysis of the sensitivity of the parameters have been omitted.

The implementation relies on a random initialization of the weight $\boldsymbol{w}$ with default ranges $(\boldsymbol{w}^h)_{ij} \in \left[\frac{-0.5}{N_h}, \frac{0.5}{N_h}\right]$ and $(\boldsymbol{w}^o)_{ij} \in \left[\frac{-0.5}{N_o}, \frac{0.5}{N_o}\right]$ where N_h and N_o denote the number of hidden units and output units respectively. As the estimated (local) minima $\boldsymbol{w}_0$ in the weight space potentially relies on the initial value of $\boldsymbol{w}$, a simple regularization has been proposed. This consists of performing three training repetitions with different random weight initializations and use the mean outputs as the posteriors.

4.3 Decision Framework

In section 2.1 a variety of popular betting options on the full time results has been presented, covering different classes of multiple bets. For simplicity and transparency only single bets will be considered from this point. In addition operations with only the simplest possible bets strengthens the applicability of the model as the model may be expanded to cover multiple bets. Provided that the involved matches are independent the probability can be determined by simply multiplying the partial probabilities. Accordingly, section 2.1 can be regarded as providing insight into the potential model expansions.

[1] `http://cogsys.imm.dtu.dk/toolbox/ann/`

The input features to the NN consists of the bookmakers' odds from a specified set of bookmakers. As discussed in section 3.2 3 convenient sets of bookmakers will be considered of $N \in \{5, 9, 37\}$ bookmakers. This means that each input sample has dimension $\boldsymbol{x}_k \in \mathbb{R}^{3 \cdot N}$, $k = 1, \ldots, M$. The input features have been standardized prior to the application of the NN, i.e. $(\tilde{\boldsymbol{x}_k})_i = \frac{(\boldsymbol{x}_k)_i - \overline{x}_i}{\sigma_i}$, where $\overline{x}_i$ and σ_i are the mean value and standard deviation, respectively, of the ith input feature on the training set.

The posteriors emitted by the NN are subsequently passed on to a decision framework that determines which selections to bet on and how much to bet. It is assumed that all output samples and hence matches are independent. The possibility of *at most* two selections per match is modelled as triple selections are assumed to be unprofitable by nature. The decision framework relies on estimates of the standardized expected return (SER) per match i,

$$\frac{\mathrm{E}\left[\pi_i\right]}{\sqrt{\mathrm{Var}\left[\pi_i\right]}} \tag{4.13}$$

where π_i is the return on the match, given a set of engaged selections. The standardization is reasonable, as the expected return is penalized by the the uncertainty of the investment. In finance this uncertainty is referred to as the volatility – a measure of price variation over time for a given asset – which in this context is defined as the standard deviation of the return. Obviously the performance of the approach relies heavily on how well-calibrated the NN is. This question will be covered in Chapter 5.

Consider a match $i \in \{1, \ldots, M\}$ and assume that a single bet has been made on one of the three selections $j \in \{1, 2, 3\}$, of size b_{ij}. The expected return on match i is defined as (see Appendix B.1)

$$\mathrm{E}\left[\pi_i\right] = (\hat{o}_{ij} p_{ij} - 1) b_{ij} \tag{4.14}$$

where $\hat{o}_{ij} = \max_{k \in \{1, \ldots, N_t\}} o_{ij}^k$ is the highest offered odds among all $N_t = 51$ bookmakers. The variance of the return on single selections is (see Appendix B.1)

$$\mathrm{Var}\left[\pi_i\right] = b_{ij}^2 \hat{o}_{ij}^2 p_{ij} (1 - p_{ij}) \tag{4.15}$$

Assuming $\hat{o_{ij}} = \frac{1}{p_{ij}}$ Eq. (4.15) reduces to $b_{ij}^2 \frac{1 - p_{ij}}{p_{ij}}$. Hence a small probability p_{ij} will have a large variance and a large probability p_{ij} a small variance. This yields the SER in match i on a singly engaged selection

$$\frac{\mathrm{E}\left[\pi_i\right]}{\sqrt{\mathrm{Var}\left[\pi_i\right]}} = \frac{\hat{o}_{ij}^2 p_{ij} b_{ij} - b_{ij}}{\sqrt{b_{ij}^2 \hat{o}_{ij}^2 p_{ij} (1 - p_{ij})}} = \frac{\hat{o}_{ij} p_{ij} - 1}{\hat{o}_{ij} \sqrt{p_{ij} (1 - p_{ij})}} \tag{4.16}$$

Note that the ratio is invariant to the bet size.

The framework is now expanded to include the possibility of betting on two selections in a single match. Suppose that two selections $m, n \in \{1, 2, 3\}$, $m \neq n$ are made. For simplicity it will be assumed that the same amount, b_i, is waged on each selection. The expected return is then (see Appendix B.1)

$$\mathrm{E}\left[\pi_i\right] = \mathrm{E}\left[\pi_{im}\right] + \mathrm{E}\left[\pi_{in}\right] = b_i\left(\hat{o}_{im}p_{im} + \hat{o}_{in}p_{in} - 2\right) \tag{4.17}$$

and the variance is (see Appendix B.1)

$$\mathrm{Var}\left[\pi_i\right] = \mathrm{Var}\left[\pi_{im}\right] + \mathrm{Var}\left[\pi_{in}\right] - 2b^2\hat{o}_{im}\hat{o}_{in}p_{im}p_{in} \tag{4.18}$$

where π_{im} and π_{in} denote the return on selections m and n respectively. The SER on two selections in match i is then

$$\frac{\mathrm{E}\left[\pi_i\right]}{\sqrt{\mathrm{Var}\left[\pi_i\right]}} = \frac{b_i\left(\hat{o}_{im}p_{im} + \hat{o}_{in}p_{in} - 2\right)}{\sqrt{\mathrm{Var}\left[\pi_{im}\right] + \mathrm{Var}\left[\pi_{in}\right] - 2b_i^2\hat{o}_{im}\hat{o}_{in}p_{im}p_{in}}} \tag{4.19}$$

$$= \frac{b_i\left(\hat{o}_{im}p_{im} + \hat{o}_{in}p_{in} - 2\right)}{\sqrt{b_i^2\hat{o}_{im}^2p_{im}(1-p_{im}) + b_i^2\hat{o}_{in}^2p_{in}(1-p_{in}) - 2b_i^2\hat{o}_{im}\hat{o}_{in}p_{im}p_{in}}} \tag{4.20}$$

$$= \frac{\hat{o}_{im}p_{im} + \hat{o}_{in}p_{in} - 2}{\sqrt{\hat{o}_{im}^2p_{im}(1-p_{im}) + \hat{o}_{in}^2p_{in}(1-p_{in}) - 2\hat{o}_{im}\hat{o}_{in}p_{im}p_{in}}} \tag{4.21}$$

The expressions obtained on single selections and double selections are now combined into a single decision criterion formulation. Let $\pi_i(s)$ denote the return on match i on a given selection combination $s \in S$, where

$$S = \{1, 2, 3, \{1, 2\}, \{2, 3\}, \{1, 3\}\}$$

is the set of all selection combinations disregarding the full selection. A decision criterion can then be formulated as

$$D_i = \begin{cases} \arg\max\limits_{s \in S} r_i(s) & \text{if} \quad \max\limits_{s \in S} r_i(s) > \tau \\ \emptyset & \text{otherwise} \end{cases} \tag{4.22}$$

where $\tau \geq 0$ is a specified threshold value, and

$$r_i(s) = \frac{\mathrm{E}\left[\pi_i(s)\right]}{\sqrt{\mathrm{Var}\left[\pi_i(s)\right]}}, \tag{4.23}$$

and $D_i \in S$ denotes the selections to bet on in the ith match. Eq. (4.23) is determined by Eqs. (4.16) and (4.21) for single and double selections respectively. The size of the bet on the engaged selections should be reflected by the size of

the SER. The higher SER, the more attractive relation between the potential return and the risk. A simple solution is to let the size be proportional to the SER, specifically equal to the SER, i.e.

$$b_i = r_i(D_i) \tag{4.24}$$

where b_i is the bet size on all selections in the set D_i. This of course assumes $D_i \neq \emptyset$, otherwise $b_i = 0$.

The proposed decision framework has some conceptual resemblance to Markowitz' efficient frontier, although the framework is far simpler. The frontier, cf. section 1.5, represents the optimal trade-off between the expected return and the risk, defined as the standard deviation of the return for a given set of assets. Similarly the SER approach penalizes selections with high risk, cf. Eq. (4.23), where the individual matches can be regarded as a portfolio with three assets/selections.

Chapter 5

Model Evaluation and Revision

The quality of the model has been evaluated by considering two aspects. Firstly the probabilistic accuracy of the model is examined, whereby it can be deduced how well the data variation in bookmakers' odds can be used as features in a forecast model. Secondly it is demonstrated how profitable the model is, as a punter in the football odds market. Based on these results extensions of the model cf. section 4.3 are proposed and tested.

5.1 Input Feature Selection

As discussed in section 3.3.4 the inherent structure of the odds predominantly originates from a two dimensional feature space. It was deduced that the bookmakers generally agree on the odds setting although minor disparities are present. Accordingly, a relevant issue is, how many bookmakers that are sufficient as input features to cover the data variation.

Three different feature sets S_5, S_9, S_{37} have been proposed containing the $N = 5, 9, 37$ most recorded bookmakers, cf. Table 3.1. This gives an input feature dimensionality of 15, 27 or 111, respectively. It is noted that $S_5 \subset S_9 \subset S_{37}$,

whereby the large feature sets contain at least the same data variation as the smallest set. The performance has been evaluated in a 10 fold cross-validation on seasons 08/09-12/13 in both leagues. The evaluation is limited to the latest five seasons because it is the largest joint set of odds between the feature sets S_5, S_9, and S_{37}, and since the latest seasons are obviously of most interest.

The NN may be regularized in many ways e.g. by the number of hidden units, adjustment of the decay parameters and by early stopping criteria. The analysis is here restricted only to vary the number of hidden units to the proposed input feature set sizes. The decay parameters are set as the default values in the applied NN toolbox. It should be emphasized that the application of cross-validation may be considered an evasion of the of the most fundamental principles in machine learning, as future (test) data is potentially used to tune the current model. However, as shall be demonstrated, the number of hidden units and the number of bookmakers hardly change the performance. A thorough sensitivity analysis of the regularization parameters has been omitted, as it is considered of little importance due to low input feature complexity, cf. section 3.3.4.

The overall purpose of the model is to properly estimate the expected return on different selection combinations in each match, and accordingly it is essential to consider the accuracy of all three class posteriors. Forecasts of the match results are of less importance, and so the classification error is an insufficient measure of accuracy. Instead the model has been evaluated in terms of the Brier score, as it encapsulates the accuracy of all three posteriors. For a three of more classes the Brier score is defined as [42, p. 1]

$$E_{\text{Brier}} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{C}(p_{ij} - \delta_{ij})^2 \tag{5.1}$$

where M is the number of samples, $C = 3$ is the number of classes, p_{ij} is the posterior of class j in match i an δ_{ij} indicates whether the event occurred in class j (1) or not (0). Table 5.1 summarizes the mean Brier scores.

Evidently the score is about 0.02 lower on the LL dataset regardless of the configuration, which suggests that the model is better calibrated on the Spanish league. The Brier score is almost invariant to the number of hidden units, suggesting a low complexity in the input features, as discussed in section 3.3.4, although $N = 37$ with one hidden units shows slight under-fitting. It also indicates that the default weight decay parameters in the neural network toolbox successfully regularize the complex models to reduce over-fitting.

In the proceeding analysis only $N = 9$ bookmakers are used, covering seasons 05/06-12/13. This choice is based on several motives. Firstly the earliest seasons

are of least interest, and so the advantage of the temporal broadness of $N = 5$ diminishes. Secondly $N = 9$ is at least as robust as $N = 5$, making $N = 9$ preferable. Thirdly $N = 37$ is computationally expensive, and does not improve performance. The number of hidden units is prospectively set at 6, although there is no statistical evidence for a superior number of hidden units.

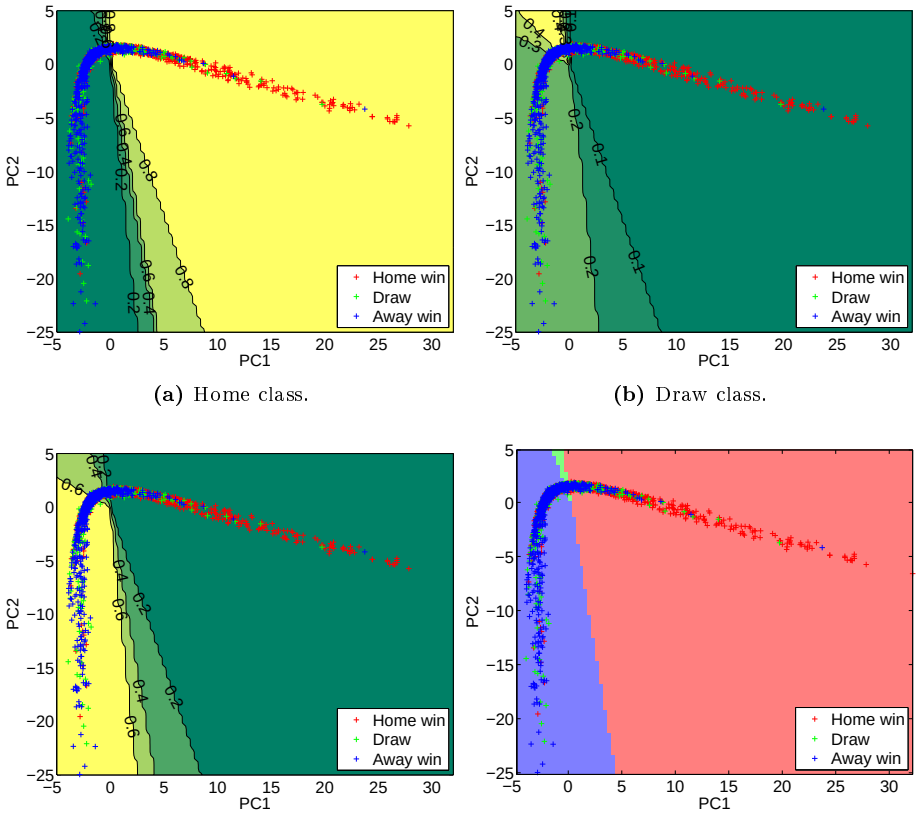
League	*# hidden units*			
	$N = 5$			
	1	3	5	10
Premier League	0.577 (0.016)	0.577 (0.014)	0.577 (0.015)	0.577 (0.014)
La Liga	0.554 (0.026)	0.553 (0.025)	0.553 (0.025)	0.553 (0.025)
	$N = 9$			
	1	6	9	18
Premier League	0.576 (0.017)	0.577 (0.018)	0.577 (0.019)	0.576 (0.019)
La Liga	0.555 (0.017)	0.553 (0.016)	0.553 (0.017)	0.554 (0.017)
	$N = 37$			
	1	10	20	40
Premier League	0.585 (0.024)	0.576 (0.029)	0.575 (0.031)	0.576 (0.028)
La Liga	0.563 (0.021)	0.554 (0.023)	0.554 (0.026)	0.553 (0.025)

Table 5.1: Mean brier score in a 10-fold cross validation on seasons 08/09-12/13 with varying number of hidden units. The brackets hold the standard deviation of the scores. The applied seasons are the largest subset of seasons containing $N = 5, 9, 37$ bookmakers.

5.2 Decision Boundaries

The complexities of the posterior class distributions have been assessed by projecting the data and distributions onto the first two principal components of the input feature space. Figures 5.1a, 5.1b and 5.1c depict the posterior distributions when training a NN on all data in the LL set, seasons 05/06-12/13 with 9 bookmakers. Based on these distributions the decision boundary between each class has been deduced by considering the most likely of the three classes, see figure 5.1d. Evidently the decision boundary between the home and away class separates the two end-segments of the characteristic U-shaped data cloud. The separation line is almost linear, and since a PCA basically is a linear transformation of standardized input features this indicates that a linear model, such as a Generalized Linear Model, may solve the classification problem equally well

as the NN. However, emphasis should be put on the distribution of the class posteriors, as outlined in section 5.1. The contour lines reveal that the distributions are complex in the high density regions (the bend of the U-shape), which justifies the use of a non-linear model. The distribution on the draw class is generally low in the relevant PC space range and grows in the most opposite direction of the home and away classes, identified as the two end segments of the U-shape. The draw class only dominates in the central part of the data cloud, where all 3 class posteriors are similar.



(a) Home class. **(b)** Draw class.

(c) Away class. **(d)** Decision boundary.

Figure 5.1: Contours of class-conditional probabilities (a+b+c) and decision boundary (d) on the La Liga dataset, seasons 05/06-12/13 with 9 bookmakers' odds.

5.3 Simulations

5.3.1 Test configuration

The profitability has been evaluated separately on the PL (Premier league) and LL (La Liga) datasets by simulating a realistic betting scenario in which the model have to decide when to bet and how much to wage. In the test set-up an initial stock of zero coins is assumed, and the player is assumed to have infinite wealth so that any negative returns can be managed. For simplicity the currency is unit-less. The development of the stock, expressed as the cumulated return, is subsequently monitored as the model completes the betting scenarios. The threshold value τ, cf. Eq. (4.22), is varied at $\tau = 0, 0.05, 0.1, 0.2$, reflecting different levels of strictness, with $\tau = 0.2$ being the most conservative model.

As discussed in section 5.1, 9 bookmakers' odds will be used as input features. This restricts the seasons to 05/06-12/13, cf. figure 3.6, which is a reasonable time horizon as earlier seasons are considered outdated. The data has been chronologically ordered on both leagues and prospective missing odds are handled by replacing the missing books with the mean book values from the remaining bookmakers.

The model is repeatedly presented with a batch of test matches corresponding to the average number of matches per week. Table 5.2 shows the duration of the 3 latest seasons of each league[1], yielding an average duration of approximately 274 days. This gives a test batch size of $7\frac{\text{days}}{\text{week}} \cdot \frac{380 \text{ matches}}{274 \text{ days}} = 9.7 \approx 10\frac{\text{matches}}{\text{week}}$.

League, season	*Start-end dates*	*Days*
PL, 10/11	14/08/10 - 22/05/11	282
PL 11/12	13/08/11 - 13/05/12	275
PL, 12/13	18/08/12 - 19/05/13	275
LL, 10/11	28/08/10 - 21/05/11	267
LL, 11/12	27/08/11 - 13/05/12	261
LL, 12/13	18/08/12 - 01/06/13	288

Table 5.2: Duration of the three recent seasons in the Premier League (PL) and the La Liga (LL), including the start and end dates.

Since the characteristics of the odds may be season dependent, a shifting training set has been proposed. The training set consists of a moving frame, which encapsulates the most recent $2 \cdot 380 = 760$ matches, corresponding to the information from two seasons. Whenever the model has processed a test batch, the

[1] en.wikipedia.org/wiki/Premier_League, en.wikipedia.org/wiki/La_liga

training frame shifts matches to encapsulate the most recent batch and thereby disregards the oldest batch. Since the training set comprises 760 matches, this set-up effectively means that the model is an 'active punter' in seasons 07/08-12/13.

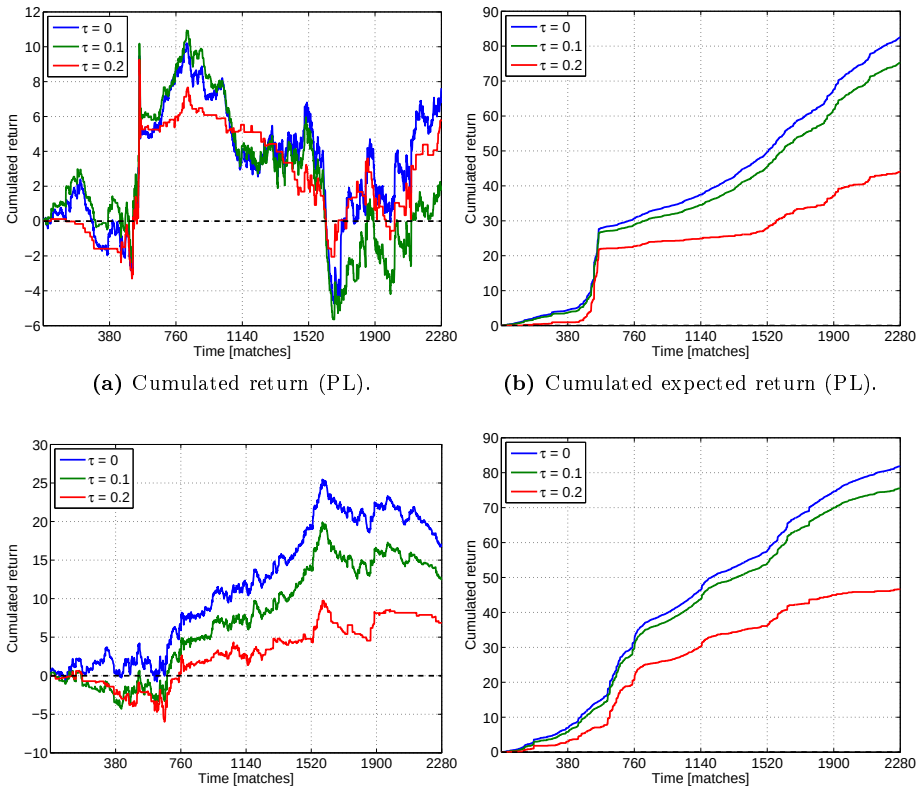
The profitability may be assessed in may ways. The most central measure is the rate of return (ROR) defined as $\frac{\text{total return}}{\text{total stake}}$, as it captures the *profit* relative to the *risk* involved. For consistency the rate is defined to be zero if no bets are made. Another important supplementary measure is the number of engaged matches (bet fraction), as it indicates the level of applicability of the model and the robustness of the profit. For instance a set-up that yields a high ROR with few bets is hardly applicable and contains high variance in terms of performance.

5.3.2 Basic model

In the first simulations the set-up described in Chapter 4 has been directly applied. This shall be referred to as the basic model, as forthcoming evidence will encourage the use of different variations of the basic model. Figure 5.2 depicts the cumulated return and expected return for the basic model. Statistics regarding the simulation are summarized in section 5.3.5, and will be discussed later.

Evidently the model perform poorly on the PL with no clear trends. Although the model encounter a series of significantly profitable matches in seasons 08/09, which is also reflected in the cumulated expected return, these returns are balanced out by a series of unprofitable matches in season 11/12. The fluctuations are less distinct on the most conservative set-up $\tau = 0.2$. On the LL dataset the cumulated return is steadily increasing in seasons 08/09-10/11 and slightly decreasing in seasons 11/12-12/13 with a global peak in season 11/12. With increasing τ the total return is reduced, as obviously fewer bets are made although the same tendency is observed.

An interesting observation is found in the cumulative expected return, as the expected return significantly exceeds the actual return. Naturally the cumulative expected return is monotonically increasing, as bets are only made on selections with non-negative expected return. The expected return and actual return should statistically, however, agree provided that the model is well-calibrated. To access how well-calibrated the model is, the estimated probabilities are compared to the actual outcome of the matches. For each class (home, draw and away) the posterior probability from the model of selection j in match i is held up against a binary function δ_{ij} indicating whether the jth selection came true or not, cf. Eq. (3.13). The samples are constructed as a coordinates with the



(a) Cumulated return (PL). **(b)** Cumulated expected return (PL).

(c) Cumulated return (LL). **(d)** Cumulated expected return (LL).

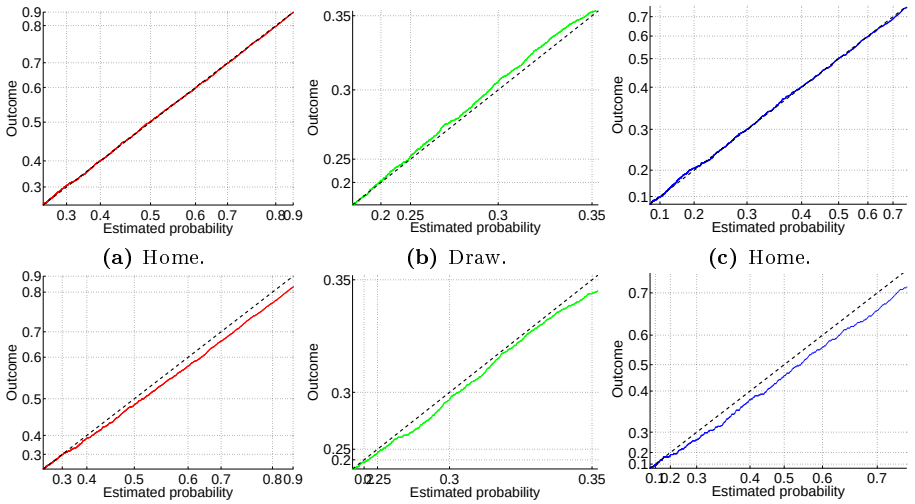
Figure 5.2: Basic model: Cumulated return and expected return on the Premier League (PL) and La Liga (LL) datasets, seasons 07/08-12/13. The vertical grid lines refer to the transition between seasons.

accumulated estimated probabilities and cumulated empirical probabilities

$$\left(\sum_{i=1}^{k} p_{ij}, \sum_{i=1}^{k} \delta_{ij}\right), \qquad k = 1, \ldots, M, \quad j = 1, 2, 3 \tag{5.2}$$

where the matches are sorted according to ascending p_{ij}. Figures 5.3 and 5.4 depict the cumulated probabilities on each selection on the PL and LL datasets respectively, with $\tau = 0$. The axes have been appropriately modified to state the probabilities, rather that the cumulative probabilities. Evidently the model is well-calibrated, when considering all selections, cf. figures 5.3a, 5.3b, 5.3c, 5.4a, 5.4b, and 5.4c. However, on engaged selections the opposite is true, cf. figures 5.3d, 5.3e, 5.3f, 5.4d, 5.4e, and 5.4f, on which the posteriors are generally

overestimated. This tendency towards overstating the posterior probabilities on selections is an inherent bias in the decision procedure. The following example elaborates the issue.



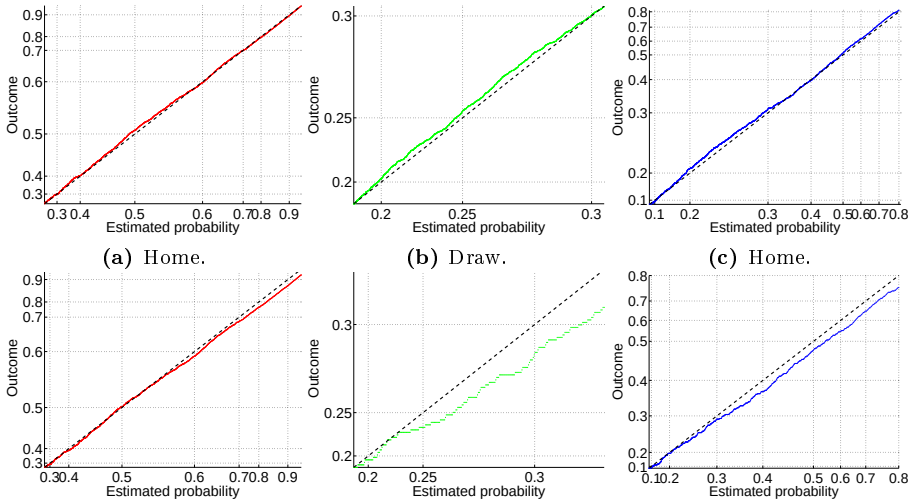
(a) Home. **(b)** Draw. **(c)** Home.

(d) Home. **(e)** Draw. **(f)** Home.

Figure 5.3: Basic model with $\tau = 0$, seasons 07/08-12/13: Comparison of cumulative outcomes and posteriors in the Premier League on all selections (a+b+c) and engaged selections (d+e+f).

Without loss of generality let $\tau = 0$ and consider a series of posteriors (p_i), $i = 1, \ldots, M'$ on a series of events with identical odds value o on a given selection type. Assume that p_i, $\forall i = 1, \ldots, M'$ are realizations from a mutual Gaussian with mean $\overline{p} = \frac{1}{o}$ and unknown variance σ^2, i.e $P \sim N(\overline{p}, \sigma^2)$. Since the model is well-calibrated, $\overline{p}$ can be assumed to be the true class probability. In order to bet on event i the model must yield a positive expected return which requires that $p_i > \overline{p}$. Denote $Q = \{p_i | p > \frac{1}{o}\}$ the discrete set of realizations, which are bet on, with $M'_Q \in \mathbb{N}$ number of elements, where $M' \geq M'_Q > 0$. Due to the properties of the normal distribution it is guaranteed that $Q \neq \emptyset$ for sufficiently large M'. Clearly then, the mean value of Q exceeds $\overline{p}$. Accordingly, the model performs a series of M'_Q bets on selections, where the expected return is estimated as positive, although the true return is in fact zero.

The property can be generalized to include cases where $\overline{p} \leq \frac{1}{o}$, i.e. where (p_i), $i = 1, \ldots, M'$ constitutes any series of selections with non-positive expected returns originating from a mutual Gaussian $P \sim N(\overline{p}, \sigma^2)$. The properties of the normal distribution will again guarantee $Q \neq \emptyset$ for sufficiently large M'.



(a) Home. **(b)** Draw. **(c)** Home.

(d) Home. **(e)** Draw. **(f)** Home.

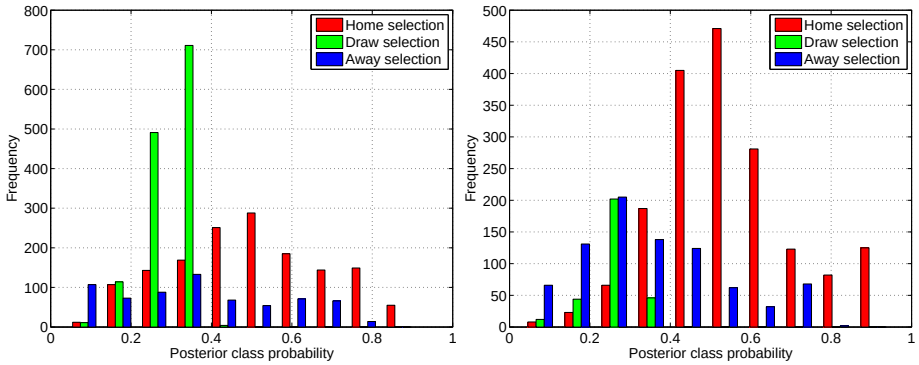
Figure 5.4: Basic model with $\tau = 0$, seasons 07/08-12/13: Comparison of cumulative outcomes and posteriors in the La Liga on all selections (a+b+c) and engaged selections (d+e+f).

Hence the mean value of the set will exceed $\frac{1}{o}$, and since $\bar{p} \leq \frac{1}{o}$ the posteriors on engaged selections will be overestimated.

Without loss of consistency this property can be extended to cover three selection events with home, draw and away outcomes. Provided that the model is presented with sufficiently many matches M this means that the posteriors are generally overstated on engaged selections, whereby the expected return generally exceeds the actual return on engaged selection. For convenience, this bias shall be referred to as a *selection bias*.

Figure 5.5 illustrates the distributions of engaged selections according to posteriors. Seemingly the distributions of the home and away selections consist of two bell-shaped components. The major component has an average value around 0.5 on the home selection and around 0.4 on the away selection. The minor component has an average of about $0.7-0.8$. This component indicates that the model is capable of detecting and subsequently betting on market inefficient odds affected by the favourite/long-shot bias. It is noted that away favourites have generally lower posteriors than home favourites, presumably due to the home ground advantage. The distinction of the components is particularly clear on the LL dataset, suggesting a more pronounced favourite/long-shot bias. The draw

selection has generally low posteriors on engaged selections, peaking around 0.3. This is consistent with figure 5.1b. Evidently the model includes more draw selections in the PL dataset than in the LL dataset.



(a) Premier League. **(b)** La Liga.

Figure 5.5: Basic model with $\tau = 0$, seasons 07/08-12/13: Posterior distributions.

5.3.3 Ensemble model

In an attempt to remove the selection bias a bootstrap aggregation of neural networks has been proposed. The basic idea of this ensemble model is to apply the lowest posteriors from a committee of networks on the engaged selections to reduce the risk of overestimated posteriors. A total of 19 bootstrap training sets with $M_{\text{train}} = 760$ samples are applied. It is noted that the new training sets on average contain

$$1 - \left(1 - \frac{1}{M_{\text{train}}}\right)^{M_{\text{train}}} \approx 0.632 = 63.2\% \tag{5.3}$$

of the samples in the original training set [43, p. 188]. This yields 20 training sets (including the full training set), which have been used to train 20 separate NN. The decision criterion for the basic model, cf. Eq. (4.22), has been modified as follows

$$D_i' = \begin{cases} \arg\max\limits_{s \in S} r_i'(s) & \text{if} \quad \max\limits_{s \in S} r_i'(s) > \tau \\ \emptyset & \text{otherwise} \end{cases} \tag{5.4}$$

where

$$r_i'(s) = \min_{j \in \{1,2,\ldots,20\}} \left(\frac{\mathrm{E}\left[\pi_{ij}(s)\right]}{\sqrt{\mathrm{Var}\left[\pi_{ij}(s)\right]}} \right) \tag{5.5}$$

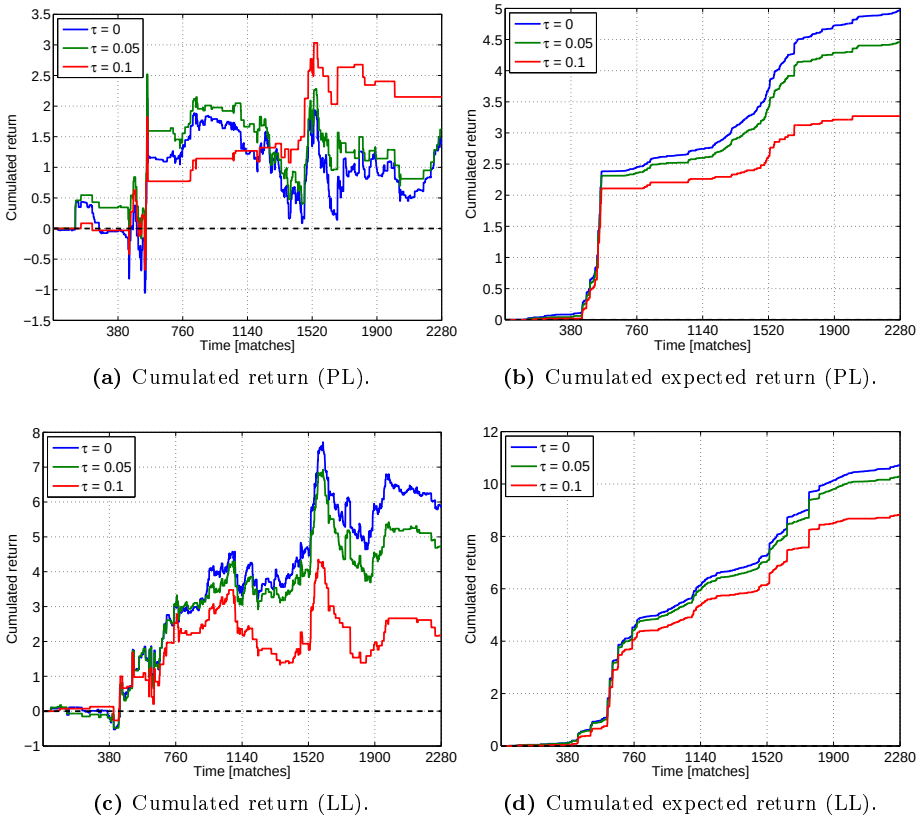
where $\pi_{ij}(s)$ is the return in match i, with a unit bet on each selection in s, when applying the posterior estimates from model j in the ensemble. S is the set of all allowable selection combinations, $S = \{1, 2, 3, \{1, 2\}, \{2, 3\}, \{1, 3\}\}$ as discussed in section 4.3. The bet size b_i on selections specified by s in match i is the same as with the basic model, i.e. $b_i = r_i'(s)$, cf. section 4.3. Consequently, the ensemble model applies a 'max-min' criterion on each match, which effectively should remove any overstated posteriors on engaged selections.

Figure 5.6 depicts the cumulated return and expected return with the ensemble model. Statistics regarding the simulations and comparisons of the models are stated in section 5.3.5. The total return developments contain many similar patterns to the basic model, although the number of bets are significantly reduced, due to the more conservative 'max-min' strategy, cf. *Bet fraction* in Tables 5.3 and 5.5. It is observed that both the cumulated return and expected return on the PL dataset increase significantly in season 08/09, suggesting a series of matches where the bookmakers generally fail to produce market efficient books. Similar effect is observed with the basic model.

Figures 5.7 and 5.8 depict the cumulated probabilities with $\tau = 0$, cf. Eg (5.2), on each selection in the PL and LL datasets respectively. The black lines state the cumulated average posteriors of the ensemble with errorbars denoting the lower and upper limits of these posterior estimates. It is noted that the 'max-min' strategy yields generally lower posteriors than the mean posteriors of the ensemble. This is reasonable, as the 'max-min' strategy is applied separately on each match i, whereby the model is allowed to choose the highest value of $r_i'(s)$, $s \in S$ from the ensemble.

It is noted that the draw and away selections in the PL dataset are generally underestimated, suggesting that the 'max-min' approach is too conservative in terms of determining the posteriors. This means that the model may overlook valuable selections that statistically will lead to profit. Contrary posteriors on the home selection are still affected by the selection bias. A likely explanation is the nature of the 'max-min' approach as the *combined* posteriors on the *combination* of engaged selections per match are minimized, cf. Eg. (5.2). Accordingly, there is no guarantee that all posteriors per match are minimized. In summary the ensemble model have difficulties with the PL dataset as the model does not seem to be perfectly well-calibrated on either selections. This might explain why the model performs poorly in the PL dataset, cf. figure 5.6a.

Unlike the PL data the ensemble model is very well-calibrated in the LL dataset and there is a reasonable similarity between the cumulated return and cumulated expected return. This leads to a better profit, cf. figure 5.6c.



(a) Cumulated return (PL).

(b) Cumulated expected return (PL).

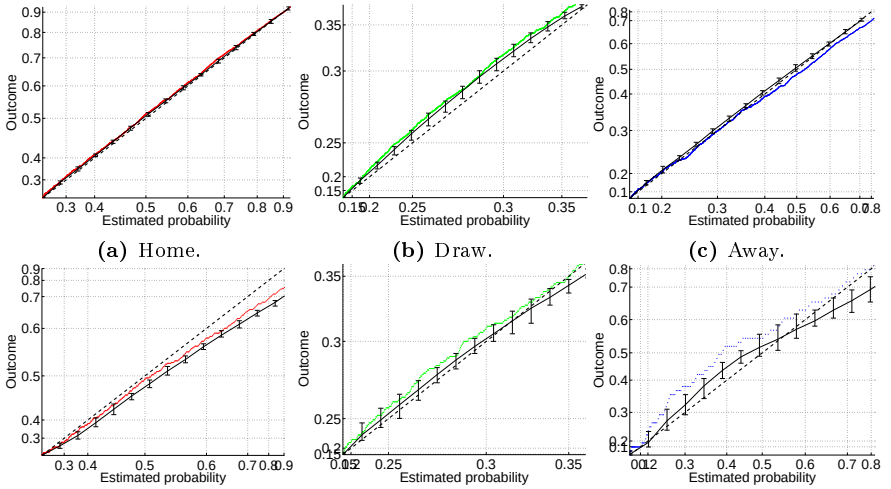
(c) Cumulated return (LL).

(d) Cumulated expected return (LL).

Figure 5.6: Ensemble model: Cumulated return and expected return on the Premier League (PL) and La Liga (LL) datasets, seasons 07/08-12/13. The vertical grid lines refer to the transition between seasons.

5.3.4 Posterior restrictive model

Consider once again the basic model. Figure 5.9 depicts the cumulated rate of return (ROR) of the selections on engaged matches with the basic strategy with $\tau = 0$. The RORs per match $i = 1, \ldots, M$ have been sorted according to ascending sum of posteriors, $p_{i,\text{sum}} = \sum_{j=D_i} p_{ij}$, on the engaged selections,



(a) Home. (b) Draw. (c) Away.

(d) Home. (e) Draw. (f) Away.

Figure 5.7: Ensemble model with $\tau = 0$, seasons 07/08-12/13: Comparison of cumulative outcomes and posteriors in the Premier League on all selections (a+b+c) and engaged selections (d+e+f).

where $D_i \in S$ encapsulates the engaged selections in the ith match using the basic model, cf. Eq. (4.22). Quite interestingly there are probabilistic regions where the ROR appears to be larger than others. Particularly on the LL dataset there is a distinctively high profitability on engaged selections with $p_{i,\text{sum}} \in [0.4, 0.5]$. Motivated by this empirical evidence of an uneven distribution of profitability potentials, an extension of the basic model that restricts $p_{i,\text{sum}}$ has been proposed.

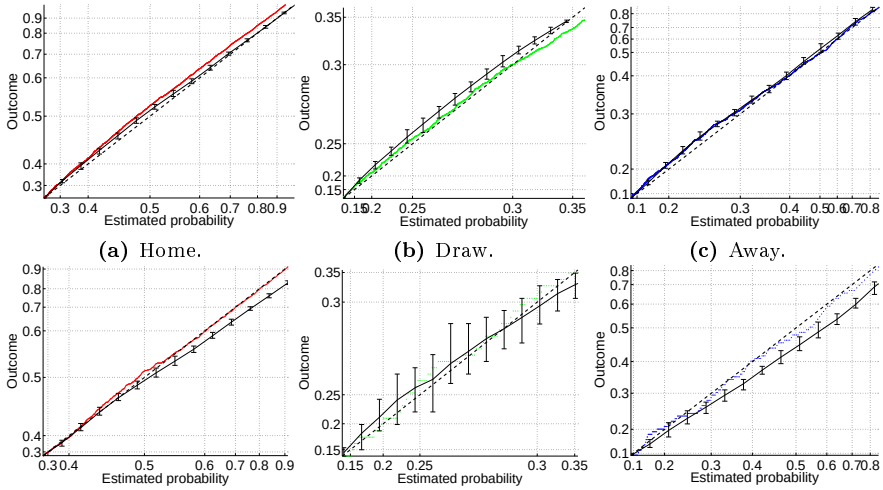
The basic idea is to apply an additional set of eight 'passive' models $\mathcal{M}_k$, $k = 1, \ldots, 8$ simultaneously with the 'active' model, each capturing mutually exclusive subsets of the matches engaged by the basic model. For a given match i, the kth passive model only allowed to pick up D_i if $p_{i,\text{sum}} \in I_k$, where I_k is the kth subintervals of $[0, 1]$, defined as

$$I_1 = [0, 0.2], \tag{5.6}$$

$$I_k = [0.1k, 0.1(k+1)[, \quad k = 2, \ldots, 7, \tag{5.7}$$

$$I_8 = [0.8, 1] \tag{5.8}$$

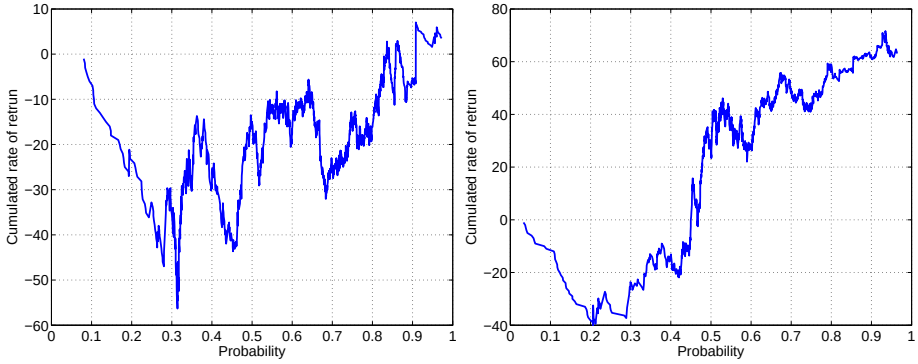
The rate of return (ROR) of each passive model of each match is added as the models are presented with new batches of test matches. For convenience the ROR equals zero if no bets are made. Similar to the basic model a limited



(a) Home. (b) Draw. (c) Away.

(d) Home. (e) Draw. (f) Away.

Figure 5.8: Ensemble model with $\tau = 0$, seasons 07/08-12/13: Comparison of cumulative outcomes and posteriors in the La Liga on all selections (a+b+c) and engaged selections (d+e+f).



(a) Premier League. (b) La Liga

Figure 5.9: Basic model with $\tau = 0$, seasons 07/08-12/13: Cumulated rate of return on engaged selections, sorted by ascending sum of posteriors per match.

memory system has been proposed, whereby RORs from only the past 760 matches are collected. Consequently, the 'passive' models are saturated only when at least 760 test matches has been encountered, which means that the

posterior restrictive model is only saturated when at least two test seasons have been processed. For a given test batch the model only engages match i if the selections D_i have $p_{i,\text{sum}}$ restricted to the interval I_{k_i}, corresponding to the 'passive' model $\mathcal{M}_{k_i}$ with highest sum of RORs across the past 760 test matches. The procedure has been formalized below.

Let R_i^k denote the cumulated ROR over the most recent 760 matches up to match i satisfying $p_{l,\text{sum}} = \sum_{j=D_l} p_{lj} \in I_k$, $\forall l = \max(1, i-760), \ldots, i-1$. R_i^k can then formulated as

$$R_i^k = \sum_{\substack{n \in \{m | p_{m,\text{sum}} \in J_k, \\ m = \max(1, i-760), \ldots, i-1\}}} \frac{\pi_n(D_n)}{B_n} \tag{5.9}$$

where B_n is the total amount betted in match n on selections $D_n \in S$ engaged by the basic model. The best subinterval I_{k_i} is then found as the interval yielding the highest cumulated ROR on the past 760 matches, i.e.

$$k_i = \underset{k \in \{1,\ldots,8\}}{\arg\max} R_i^k \tag{5.10}$$

The decision criterion in the basic model, cf. Eq. (4.22), is then extended by adding an additional layer of restrictions

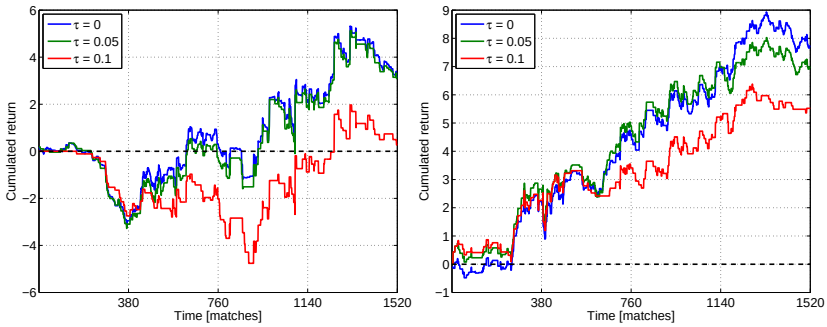
$$D_i'' = \begin{cases} D_i & \text{if } p_{i,\text{sum}} \in I_{k_i} \\ \emptyset & \text{otherwise} \end{cases} \tag{5.11}$$

where $r_i(s)$ is defined as in Eq. (4.23), D_i is the selections to bet on from the basic model, and D_i'' contains the selections to bet on with the posterior restrictive model. Consequently, the new model engages a subset of the selection engaged by the basic model.

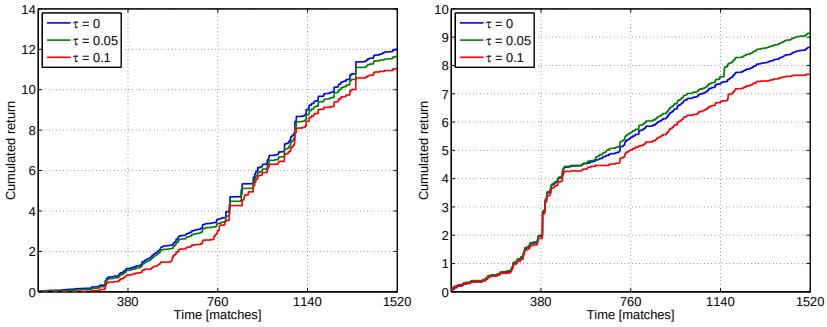
Figure 5.10 depicts the cumulated return and expected cumulated return on the PL and LL datasets, as well as the choice of I_{k_i} in each batch. The performance has only been evaluated on the latest four seasons 09/10-12/13, due to the training phase of the passive models, as previously discussed. By application of both the basic model and the ensemble model it was observed that the cumulative return reached a global peak in season 11/12 in the LL dataset, cf. figures 5.2 and 5.6. This trend is however, not present with the posterior restrictive model which profits very steadily on the LL dataset, except minor drops in for instance season 12/13. A reasonably steady increase in profit is also observed in the PL dataset during the latest three seasons. Statistics regarding the simulation are summarized in the next section. Contrary to the basic model, the restrictive model yields very similar developments of the expected return and true return.

The acceptable posterior region I_{k_i} in the PL dataset lies steadily at $[0.3, 0.4]$ in season 09/10, and $[0.4, 0.5]$ in seasons 10/11-12/13. With the LL dataset

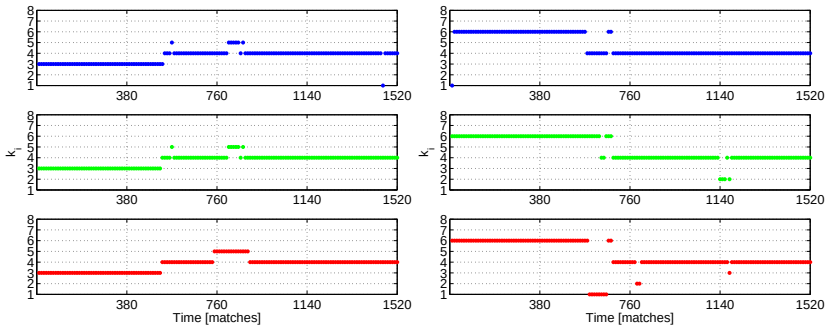
the posteriors are generally restricted to $[0.6, 0.7]$ in seasons 09/10-10/11 and $[0.4, 0.5]$ in season 11/12-12/13. This is coherent with figure 5.9b, as I_4 and I_6 are posterior regions with positive ROR on many engaged selections by the basic model.



(a) Cumulated return (PL). **(b)** Cumulated return (LL).



(c) Cumulated expected return (PL). **(d)** Cumulated expected return (LL).



(e) Development of k_i (PL). **(f)** Development of k_i (LL).

Figure 5.10: Posterior restrictive model: Cumulated return and expected return on the Premier League (PL) and La Liga (LL) datasets, seasons 09/10-12/13, and developments of posterior restrictions, cf. eq (5.10).

5.3.5 Summary of results

Tables 5.3 and 5.4 summarize key statistics of the model performance when applying on the PL dataset with varying threshold value τ. Tables 5.5 and 5.6 summarize the same statistics with application of the LL dataset. Although most of the statistics are self-explanatory some important features are described below.

5.3.5.1 Basic model, Premier League

The basic model yields minor negative profits and accordingly negative RORs, with $\tau = 0, 0.2$. Using $\tau = 0.1$ the loss magnitude is significantly higher. Regardless of τ the profit is, however, very high on the latest season 12/13. Quite remarkably the model bets on virtually all matches with $\tau = 0$, which is halved with $\tau = 0.1$ and further reduced significantly with $\tau = 0.2$.

The error rate, corresponding to the fraction of engaged matches with negative profit, is about 50 % regardless of τ. The error is partly balanced out by the high mean odds of about 3, and winning selections with high odds, cf. the maximum odds won. This is true for the ensemble model and the posterior restrictive model, as well. The selection distributions are very similar for all values of τ. Half the bets are made on home-draw doubles, followed by singles on home and draw-away doubles. Evidently betting on multiple selections yields high expected returns, which in addition reduces the vulnerability of losses.

5.3.5.2 Basic model, La Liga

The total return and accordingly the overall ROR is positive for all τ, although all configurations lose a significant amount of money on season 12/13. The bet fraction is generally higher and the error rate lower compared to the PL dataset, and possibly this combination of positive factors explains why the model gains significantly more with the LL dataset. On the other hand the mean and max odds won are significantly lower, suggesting that the model generally gains less per engaged match but wins on more bets. The model bets predominantly on single home selections, weakly followed by away singles. Thus single selections are preferred, contrary to the PL dataset. Evidently the characteristics of the betting pattern is very different between the two datasets.

5.3.5.3 Ensemble model, Premier League

The ensemble model yields minor positive total returns for all thresholds. $\tau = 0.1$ implies very few bets and accordingly the ROR is very high. The error rate with $\tau = 0, 0.05$ is similar to the basic model. With $\tau = 0.2$ the error is significantly lower, which most likely is a consequence of the high variance on the statistics, as only 2.57% of all matches have been engaged.

The max-min approach, cf. Eqs. (5.4) and (5.5), implies that the posteriors on engaged selections are generally lower than the equivalent posteriors with the basic model. Accordingly, the bet sizes, which are equal to the the standardized expected return (SER), will likewise be lower. This is evident in Table 5.4, as the bets are several factors smaller.

The clear majority of bets are made on home-draw doubles, followed by home singles. Compared with the basic model, the ensemble model cuts away many home singles and draw-away doubles, presumably because the variances of the posteriors with these combinations are substantial, whereby the SER is reduced. Consequently, fewer matches are engaged with these combinations.

5.3.5.4 Ensemble model, La Liga

Similar to the basic model the error rate is generally lower and the bet fraction is higher compared to the PL dataset. Except for $\tau_e = 0.1$, the model demonstrates similar total ROR to the basic model, although the return on season 12/13 is now positive. This indicates that the model is strongly affected by the selection bias in season 12/13. As with the basic model generally lower winning odds are observed, compared to the PL dataset, and the selections are focused on home singles followed by home-draw doubles.

5.3.5.5 Posterior restrictive model, Premier League

The posterior restrictive model produces relatively high returns as well as high ROR. However, with $\tau = 0.1$ the model losses on all engaged selections in season 09/10. The bet fraction is halved as τ is doubled. It is observed that the error rate is about 62%, regardless of τ. This is consistent with the fact that I_{k_i} predominantly assumes the intervals $[0.3, 0.4]$ and $[0.4, 0.5]$, whereby an error rate of about 60% is expected. Specifically in the first season the error rate is large, as the model chooses a low interval index k_i, cf. figure 5.10e.

The generally low values of the acceptable posteriors can be perceived as a high risk-willingness, which is supported by generally higher mean odds values, compared to the latter two models. The low posteriors also implies that relatively fewer bets are made on doubles, as these selection combinations typically have higher sums of posteriors. With $\tau = 0, 0.05$ bets are frequently made on home and draw singles. Away selections are primarily represented in doubles with draw selections. The model never bets on home-away doubles, possibly because the sum of posteriors consistently exceeds I_{k_i}, as one of the teams is considered at least a weak favourite.

5.3.5.6 Posterior restrictive model, La Liga

The ROR is significantly higher than any of the other models. Although the mean winning odds are small compared to the PL dataset, the significantly lower error rates imply very high profits. The model mainly applies the intervals $[0.4, 0.5]$ and $[0.6, 0.7]$ in equal proportions, yielding an expected success rate of about 55% which is consistent with the error rate around 47%. As the other two models, mainly home single selections are engaged.

Set-up	*Total return*	*ROR* [%]	*Bet frac.* [%]	*Error* [%]
$\tau_b = 0$	-0.938 (-4.54,1.89, -3.92,5.63)	-0.313 (-8.54,2.23, -4.82,7.00)	99.1 (99.2,98.7, 99.7,98.9)	51.8 (50.9,45.9, 59.4,50.8)
$\tau_b = 0.1$	-6.89 (-4.53,0.387, -6.11,3.37)	-3.06 (-12.9,0.565, -9.80,5.67)	53.9 (47.4,63.4, 52.4,52.6)	51.1 (50.0,43.6, 58.8,53.5)
$\tau_b = 0.2$	-0.176 (-0.468,-3.64, -1.24,5.17)	-0.293 (-4.87,-34.1, -5.60,29.5)	9.41 (9.21,6.84, 11.8,9.74)	53.1 (45.7,61.5, 62.2,43.2)
$\tau_e = 0$	0.105 (-0.0764,0.122, -0.374,0.434)	0.284 (-1.61,0.751, -3.30,9.25)	32.4 (23.7,46.3, 38.9,20.8)	50.9 (50.0,42.0, 61.5,51.9)
$\tau_e = 0.05$	0.0633 (0.109,-0.106, -0.415,0.475)	0.272 (4.07,-0.965, -5.58,22.5)	9.93 (6.32,17.4, 12.4,3.68)	45.7 (41.7,43.9, 53.2,35.7)
$\tau_e = 0.1$	1.38 (0.498,1.28, -0.143,-0.254)	15.9 (37.9,35.0, -4.17,-100)	2.57 (2.37,3.95, 3.68,0.263)	38.5 (22.2,26.7, 57.1,100)
$\tau_r = 0$	3.17 (-2.95,3.27, 1.85,1.00)	6.21 (-46.5,22.2, 10.7,7.91)	29.3 (22.9,41.6, 28.2,24.7)	62.3 (78.2,54.4, 62.6,60.6)
$\tau_r = 0.05$	3.10 (-3.10,2.84, 2.20,1.15)	6.92 (-65.3,22.3, 13.9,10.1)	16.4 (10.5,27.6, 14.5,13.2)	61.6 (85.0,54.3, 60.0,60.0)
$\tau_r = 0.1$	0.269 (-2.74,0.716, 1.15,1.15)	0.750 (-100,7.68, 7.55,13.5)	8.42 (3.95,12.1, 10.3,7.37)	65.6 (100,63.0, 61.5,57.1)

Table 5.3: Statistics on model profitability on the Premier League dataset, seasons 09/10-12/13. τ_b, τ_e and τ_r refer to the basic model, the ensemble model and the posterior restrictive model, respectively. The brackets hold the statistics per season, ordered chronologically. The error states the fraction of engaged matches with negative profit.

Set-up	*Mean bet* (*min./max.*)	*Mean odds won* (*min./max.*)	Singles [%]	Doubles [%]
$\tau_b = 0$	0.118 ($< 0.01/0.453$)	2.91 (1.12/29.0)	(18.9, 8.16, 5.77)	(46.6, 15.3, 5.91)
$\tau_b = 0.1$	0.160 (0.100/0.453)	3.01 (1.17/15.0)	(18.0, 5.24, 6.59)	(52.7, 14.5, 3.78)
$\tau_b = 0.2$	0.254 (0.201/0.453)	3.45 (1.20/15.0)	(16.8, 4.20, 14.0)	(51.7, 11.9, 1.40)
$\tau_e = 0$	0.0423 ($< 0.01/0.296$)	2.93 (1.17/29.0)	(10.8, 9.74, 4.26)	(63.7, 5.48, 7.10)
$\tau_e = 0.05$	0.0868 (0.0502/0.296)	2.90 (1.25/6.50)	(12.6, 6.62, 5.96)	(68.9, 5.30, 1.99)
$\tau_e = 0.1$	0.129 (0.101/0.296)	2.98 (1.30/6.00)	(15.4, 7.69, 10.3)	(69.2, 0.00, 0.00)
$\tau_r = 0$	0.0851 ($< 0.01/0.357$)	3.67 (1.99/10.0)	(37.7, 22.6, 5.16)	(14.8, 19.7, 0.00)
$\tau_r = 0.05$	0.128 (0.0501/0.357)	3.96 (1.99/10.0)	(37.6, 20.0, 2.40)	(23.2, 16.8, 0.00)
$\tau_r = 0.1$	0.177 (0.100/0.357)	4.60 (2.15/10.0)	(31.3, 8.59, 2.34)	(45.3, 12.5, 0.00)

Table 5.4: Odds statistics on model on the Premier League dataset, seasons 09/10-12/13. τ_b, τ_e and τ_r refer to the basic model, the ensemble model and the posterior restrictive model, respectively. 'Singles' refers to fraction of bets on selection (H, D, A) and 'Doubles' to fraction of bets on selections $(H + D, D + A, H + A)$, where H, D and A refer to home, draw, and away selections, respectively.

Set-up	*Total return*	*ROR* [%]	*Bet frac.* [%]	*Error* [%]
$\tau_b = 0$	10.9 (6.02,7.45, 2.61,-5.21)	4.28 (7.99,11.3, 4.01,-11.0)	99.8 (99.7,99.7, 100,99.7)	48.7 (45.1,40.1, 52.4,57.3)
$\tau_b = 0.1$	9.79 (4.82,6.53, 2.18,-3.75)	4.86 (7.30,12.5, 3.99,-13.1)	62.7 (74.5,62.4, 69.7,44.2)	48.3 (43.8,39.7, 55.1,57.1)
$\tau_b = 0.2$	6.52 (2.39,2.08, 3.33,-1.27)	7.76 (6.51,9.68, 15.6,-27.2)	17.0 (28.7,17.4, 17.6,4.47)	42.1 (36.7,40.9, 49.3,52.9)
$\tau_e = 0$	2.81 (0.434,1.16, 1.21,<0.01)	5.67 (2.95,9.12, 8.15,0.0288)	42.8 (41.3,42.9, 48.9,37.9)	46.2 (42.7,37.4, 50.0,54.9)
$\tau_e = 0.05$	1.63 (0.126,0.805, 0.483,0.216)	4.10 (0.992,7.67, 3.91,5.14)	20.6 (23.9,21.6, 26.3,10.5)	41.9 (34.1,37.8, 53.0,40.0)
$\tau_e = 0.1$	-0.344 (-0.157,-0.425, 0.0421,0.195)	-1.40 (-1.67,-7.32, 0.571,9.45)	9.01 (13.4,8.42, 10.5,3.68)	40.9 (33.3,43.8, 47.5,42.9)
$\tau_r = 0$	7.73 (2.22,2.38, 2.26,0.871)	16.2 (18.5,14.9, 22.7,8.83)	25.1 (22.1,28.7, 20.3,29.2)	45.7 (33.3,38.5, 55.8,55.0)
$\tau_r = 0.05$	6.99 (2.66,2.23, 2.08,0.0207)	15.1 (23.4,13.6, 21.5,0.232)	20.3 (16.1,24.5, 18.4,22.4)	46.9 (32.8,39.8, 57.1,56.5)
$\tau_r = 0.1$	5.52 (2.26,1.15, 1.69,0.426)	16.2 (23.4,10.3, 20.7,8.37)	12.2 (12.6,12.9, 13.9,9.47)	47.8 (35.4,44.9, 56.6,55.6)

Table 5.5: Statistics on model profitability on the La Liga dataset, seasons 09/10-12/13. τ_b, τ_e and τ_r refer to the basic model, the ensemble model and the posterior restrictive model, respectively. The brackets hold the statistics per season, ordered chronologically. The error states the fraction of engaged matches with negative profit.

Set-up	*Mean bet* (*min./max.*)	*Mean odds won* (*min./max.*)	Singles [%]	Doubles [%]
$\tau_b = 0$	0.134 ($< 0.01/1.19$)	2.22 (1.08/13.4)	$(65.0, 1.85, 8.24)$	$(7.25, 5.21, 12.7)$
$\tau_b = 0.1$	0.178 (0.100/1.19)	2.21 (1.08/11.5)	$(70.5, 2.31, 8.50)$	$(6.51, 5.88, 6.51)$
$\tau_b = 0.2$	0.286 (0.201/1.19)	2.02 (1.13/8.50)	$(75.7, 3.09, 7.72)$	$(3.47, 6.18, 3.86)$
$\tau_e = 0$	0.0615 ($< 0.01/0.648$)	2.23 (1.12/8.50)	$(69.5, 1.85, 5.08)$	$(12.9, 1.85, 9.08)$
$\tau_e = 0.05$	0.109 (0.0500/0.648)	2.07 (1.13/7.84)	$(79.2, 1.92, 3.19)$	$(10.5, 1.92, 3.83)$
$\tau_e = 0.1$	0.162 (0.100/0.648)	1.86 (1.13/3.50)	$(86.1, 0.730, 2.19)$	$(5.84, 1.46, 3.65)$
$\tau_r = 0$	0.105 ($< 0.01/0.513$)	2.51 (1.53/6.00)	$(64.6, 0.262, 15.2)$	$(12.1, 6.56, 1.31)$
$\tau_r = 0.05$	0.133 (0.0512/0.513)	2.33 (1.53/4.10)	$(74.1, 1.29, 11.7)$	$(8.09, 3.88, 0.971)$
$\tau_r = 0.1$	0.170 (0.101/0.513)	2.31 (1.61/3.80)	$(82.8, 0.538, 9.14)$	$(3.23, 3.76, 0.538)$

Table 5.6: Odds statistics on model on the La Liga dataset, seasons 09/10-12/13. τ_b, τ_e and τ_r refer to the basic model, the ensemble model and the posterior restrictive model, respectively. 'Singles' refers to fraction of bets on selection (H, D, A) and 'Doubles' to fraction of bets on selections $(H + D, D + A, H + A)$, where H, D A refer to home, draw, and away, respectively.

5.4 Quantile Analysis of Posteriors

A central issue is whether the probabilities *implied by the odds*, denoted $\tilde{p}_{ij}$, correspond to the *posterior probabilities* p_{ij} estimated by the model. A reasonable assumption is that $\tilde{p}_{ij}$ is captured by the reciprocal odds value normalized by the bookmakers' profit margin. Accordingly, define $\tilde{p}_{ij}$ as

$$\tilde{p}_{ij} = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{o_{ij}^{k}(\eta_i^k + 1)}, \qquad i = 1, \ldots, M, \quad j = 1, 2, 3 \tag{5.12}$$

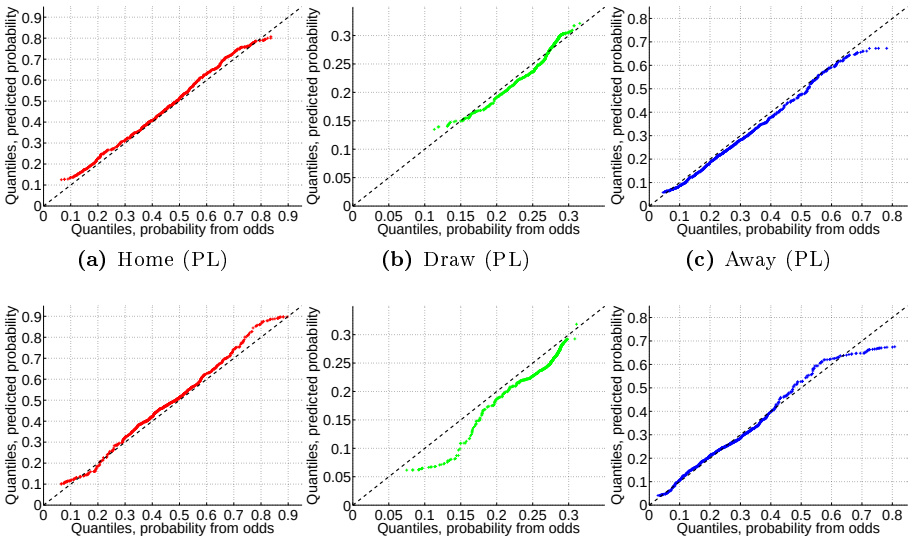
where η_i^k is the overround of bookmaker $k = 1, \ldots, N$ in match $i = 1, \ldots, M$, cf. Eq. (3.12).

The distributions of $\tilde{p}_{ij}$ and p_{ij}, $j = 1, 2, 3$ have been compared using a Q-Q plot on each selection j with the $N = 9$ most recorded bookmakers on both datasets, seasons 05/06-12/13. A hold-out validation has been carried out with $\frac{3}{4}$ of the data as training set and $\frac{1}{4}$ as test set.

Figure 5.11 depicts the Q-Q-plots comparing $\tilde{p}_{ij}$ and p_{ij}, $j = 1, 2, 3$ on the test set. The Q-Q plot is a probability plot, where the quantiles of two distributions are compared by plotting their respective quantiles against each other. If the quantiles of $\tilde{p}_{ij}$ and p_{ij} constitute a straight line $y = \alpha x$ the distributions are linearly related, specifically if $\alpha = 1$ the distributions are identical. If $\alpha < 1$ in a given quantile region R the distribution of $\tilde{p}_{ij}$ is more dispersed than the distribution of p_{ij} in R. Contrary p_{ij} is more dispersed than $\tilde{p}_{ij}$ in R, provided that $\alpha > 1$ in R.

The distributions around the median are very similar on the home and away selections, as the central quantiles seem equally spaced. This indicates that the odds indeed agree with the class probabilities on matches with roughly even opposing teams. However, the lower and upper quantiles on the two selections demonstrate that the distributions of p_{i1} and p_{i3} are far more dense than $\tilde{p}_{i1}$ and $\tilde{p}_{i3}$ respectively in those regions. Since the model is well-calibrated, cf. figures 5.3a, 5.3b, 5.3c, 5.4a, 5.4b, and 5.4c, these characteristics can be directly interpreted as biases in the odds.

On home selections the bookmakers generally understate the winning chances of clear favourites by setting the favourite odds too high and simultaneously overstate the winning chances of very weak teams by setting odds too low. This 'regularization' of odds in matches with extremely uneven teams, confirms the existence of the favourite/long-shot bias, cf. section 2.3.3. Quite interestingly the bias far more significant in the LL dataset. Presumably the Spanish punters bets more irrationally as more bets are made on highly unlikely outcomes,



(a) Home (PL) **(b)** Draw (PL) **(c)** Away (PL)

(d) Home (LL) **(e)** Draw (LL) **(f)** Away (LL)

Figure 5.11: Q-Q-plots comparing the distribution of the probabilities directly extracted from the odds with the posteriors estimated by the model. The data consists of odds data from seasons 05/06-12/13 from the 9 most recorded bookmakers in the Premier League (PL) and the La Liga (LL) datasets.

forcing the bookmakers to enhance the bias in the odds. This could be caused by e.g. strong team loyalty to small weak teams in matches where the opponent is far stronger. This conclusion is consistent with figure 5.5, from which it was also deduced that the bias is more prevalent in the Spanish league.

Although $\tilde{p}_{i3}$ is significantly more dispersed than p_{i3} in the high quantiles, the values of $\tilde{p}_{i3}$ are generally higher than p_{i3}. This contradicts with the observations on the home selections, as the winning chances of strong away favourites are generally overrated, i.e. the odds are too low. The curves indicate that the actual winning frequency of away favourites is more persistent and smaller at $p_{i3} \in [0.6, 0.7]$, than the odds appear to reveal. The tendency is particularly profound on the LL dataset. Presumably bookmakers have difficulties estimating the strength of strong away teams, leading to very conservative away odds. A further discussion of this issue is found in section 6.4. This effect completely dominates the favourite/long-shot bias.

It is noted that $\max_i p_{i1} \approx 0.9$ on the LL dataset and $\max_i p_{i1} \approx 0.82$ on the

PL dataset. Thus, whenever the home team is a clear favourite, the probability of a home win is significantly higher in the LL. This either suggests that the home team advantage is more profound in the LL with strong home teams, or that strong teams in the PL are simply relatively stronger than weak teams in the league. Indeed figure 3.10 shows that the Spanish League have two superior teams (Real Madrid and F.C. Barcelona) whereas the English league only has one (Manchester Unitied) with winning rates exceeding 70 %.

Additionally it is observed that $\max_i p_{i3} \approx 0.675$ on both datasets, which is significantly lower than $\max_i p_{i1}$. This indicates a generally higher confidence in home wins, when the home team is the favourite, compared to away wins when the away team is the favourite. This is consistent with figure 3.2b, stating that the away odds on away wins generally exceed the home odds on home wins, as a consequence of the bookmakers' substantial confidence in the home ground advantage.

The distribution of the draw selection differs from the other. There is no indication of an S-shape, as the favourite/long-shot bias does not directly affect the draw selection. The probabilities are generally small, $\max_i p_{i2} \approx 0.35 = 35\%$, on both datasets as draws are unlikely. Particularly the LL dataset contains many low posteriors ($p_{i2} < 0.1$) on the draw selection, showing that draws are more unlikely in the Spanish league. Additionally p_{i2} is generally lower than $\tilde{p}_{i2}$ in the LL set, suggesting that odds are generally unattractive on the draw selection. This is supported by Tables 5.4 and 5.6, as the fraction on bets on single draw selections are consistently much lower in the LL dataset.

Chapter 6

Discussion

6.1 Research Impact

The content of this thesis is motivated by the conclusions from former researches. It is, however, perceived that many of the results presented this thesis have not been elaborated to this extend before as generally little research is publicly available. This includes the statistical analysis of a large odds dataset, which reveals strong evidence of notorious bias phenomena within the odds setting framework in association football, causing market inefficiencies. Additionally it illuminates the existences of distinguishable odds characteristics regarding the three different outcome classes, and league characteristics.

The proposed betting model has been successfully applied to a La Liga (LL) odds dataset, demonstrating that a simple model solely based on a small class publicly accessible statistics is able to profit on the odds market. The key assumptions of this approach are that 1) the bookmakers have already done the research to accurately subjectively estimates the outcome probabilities, 2) this distribution is concealed by a 'social filter' to encompass biases and, 3) the model is capable of filtering out the biases. Accordingly, the proposed model outsources the statistical work to a crowd of external experts (bookmakers) to profit on the bookmakers themselves. Given the generalizability of the model it may be applied to other football leagues possibly with even better results,

other types of bets or even other sports, which are sufficiently covered with odds statistics.

6.2 Model Performance

The initially proposed model is based on maximizing the standardized expected return in each match by engaging an appropriate combination of outcome selections. Evidently this basic model performs very well on the La Liga (LL) data set in earlier seasons 08/09-10/11. However, on recent seasons which are of most interest the model yields a negative return. Contrary the model gives positive return on the Premier League (PL) dataset in the latest season, 12/13, and overall insignificant returns in preceding seasons. Overall the basic model is not a robust betting model, and the practical application in future seasons is questionable.

The performance measures are generally similar with the ensemble model, although it demonstrates the existence of a selections bias and how one could handle it. Figure 5.6 depicts the cumulated returns and expected returns. Although the model is reasonably calibrated, particularly on the LL dataset, cf. figures 5.8d, 5.8e and 5.8f, the expected returns still exceeds the actual return. A central contribution to the prevailing deviation is allegedly the betting sizes, cf. Eq. (4.24), as overestimated matches will always have the highest bets. Consequently the cumulated expected returns are dominated by bets, where the selection bias is not completely removed.

A highly relevant question is whether it is even necessary to apply an ensemble, as the threshold value τ may suffice as a safety margin. The selection bias may be handled sufficiently by an appropriate value of τ, as it filters out critical selections balancing between positive and negative returns, which are highly likely to be dominated by the selection bias. Although the bias would still present, the model should not be penalized by it.

The posterior restrictive model differs significantly from the latter two models, as it also applies empirical evidence, rather than acknowledged statistical methods. It utilizes league specific characteristics incorporated in the rate of return to properly adjust selection restrictions. The performance of this model exceeds the latter two with generally steady returns and clearly demonstrates practical applicability. Quite remarkably the true return and expected return are very similar with the LL dataset, cf. figures 5.10b and 5.10d. Since the model is well-calibrated in the whole posterior range, cf. figure 5.4, and not particularly in the posterior region specified by I_{k_i}, it would seem that the selection bias

is particularly low in I_{k_i}. Despite several attempts no clear evidence has been found in this regard.

According to the favourite/long-shot bias theory one would anticipate the best selections to be on favourite teams. This is, however, not the case. Generally the posterior restrictive model limits its scope of posteriors to $[0.4, 0.5]$ on the recent seasons 11/12-12/13 in both leagues, which comprises a 'gray area' between the highly likely and highly unlikely selections. It is a well-known phenomenon that humans are poor probability estimators. Possibly few people are willing to bet on these unclear selections, and the bookmakers are subsequently forced to increase the odds on those selections to obtain balanced books. Alternatively, and more likely, the selections in the 'gray area' may encourage punters to put their bets on the selections that they want to be realized rather than the statistically superior selection, which leads to the significant inefficiencies in the posterior range $[0.4, 0.5]$.

If the scope is broadened to all 4 test seasons 09/10-12/13, it is observed that the mean odds lie between 3.67-4.6 in the PL dataset and 2.32-2.51 in the LL dataset, cf. Tables 5.4 and 5.6, depending on the threshold value. The mean odds are generally higher in the English leagues as the acceptable interval I_{k_i} generally restricts the model to selections with lower posteriors and consequently higher odds. This means that the selections in the PL dataset are generally biased *against* the punter due to the favourite/long-shot bias. Even so the model still profits on the latest 3 seasons in the English league, and if the bias was not present the profit would possibly be even higher.

In sections 1.3.5 and 1.3.4 the profitability capacities of other betting models have been outlined and may be used for comparison. Ref. [24] proposed an expert constructed Bayesian network and used a standard (unspecified) profitability measure with varying discrepancy levels as betting criterion on the PL season 10/11. Among other statistics the performance was measured by the overall profit/loss ratio, which is identical to the rate of return (ROR) measurement in Table 5.3. Therefore the models can be directly compared in season 10/11. By application of the posterior restrictive model with $\tau_r = 0, 0.05, 0.1$ the RORs are 22.2%, 22.3% and 7.68%, respectively, and the bet fractions are 41.6%, 27.6% and 12.1%, respectively. Similar bet fractions in [24] are found using discrepancy levels $\geq 5\%$, $\geq 7\%$ and $\geq 10\%$ yielding RORs equal to 8.4%, 12.1% and 20.4%, cf. Table 1.1. Apart from the most conservative set-up with $\tau_r = 0.1$ the posterior restrictive model is at least on par with the Bayesian network model. It is considered unsuitable to derive more precise conclusions due to the relatively low number of engaged matches.

Ref. [25] applied a NN model and tested it on 400 random matches from the LL seasons 07/08-10/11 in 5 runs. The model yielded an average profit of 0.16

per unit bet which corresponds to a ROR of 16 %. A direct comparison is not possible since the samples are mixed from different seasons and chronologically unordered. In seasons 09/10 and 10/11 the most profitable posterior restrictive model ($\tau_r = 0$) yields RORs equal to 22.1% and 28.7%. These seasons may coincide with the test samples in [25] and indicate that the posterior restrictive model is at least on par with the profitability capacity of the model in [25].

6.3 League Characteristics

The odds analysis and model simulations provide statistical characteristics of the two considered leagues and odds datasets. Although both leagues are affected by several biases consisting of the home ground advantage and the favourite/long-shot bias, with team bias being a special case, the biases are more prominent in the LL dataset, cf. figure 5.11. Since the odds reflect the public opinion, this observation indicates that punters on the Spanish league bet less rationally and more 'by heart'. Specifically Spanish fans may display higher team loyalty to weak teams. In contrast, evidence suggests that the Premier League is more irregular than La Liga in terms of match outcomes. Figures 3.4 and 3.5 indicates a more 'noisy' PL dataset with less distinct separation between the classes. In addition the Brier score, cf. section 5.1 is higher on the PL dataset, inevitably leading to generally poorer forecast capabilities, cf. section 5.3. Even though it has been demonstrated that PL odds are slightly more attractive, cf. Table 3.4, this clearly does not sufficiently compensate for the outcome irregularities. In order to better account for the irregularities the information span in the input features must be expanded by incorporating e.g. expert knowledge as proposed by Refs. [23] and [24] such as quantifiable measures of team spirit, key player's form, fatigue, etc.

6.4 Odds Characteristics

In section 5.1 it was demonstrated that the predictive precision of the neural network is highly indifferent to whether 5, 9 or 37 bookmakers' odds are used as input features. The accuracy improvement of using a committee of models (bookmakers) strongly depends on the correlation between the models, as high correlation implies generally small accuracy improvements, cf. [22, p. 657]. Since strong correlation has been observed between the bookmakers in individual matches, cf. figure 3.7, the performance improvements are insignificant to whether 5, 9 or 37 bookmakers are used, as few bookmakers sufficiently cover

the data information of all bookmakers. This observation suggests that the data consists of two major components. The first component is the highly dominating low frequency features, adequately captured by few bookmakers. The other component consists of high frequency features, encapsulating the individual bookmakers' characteristics, which the model mainly recognizes as white noise. From this perspective the initial idea of applying a large crowd of experts (bookmakers), seems unnecessary as the disagreements between the bookmakers hardly contribute to the model's forecast capability. Most likely the fluctuations are simply the result of each bookmaker's attempts to balance its books. This is consistent with figures 3.4 and 3.5 which demonstrate low data complexity. Indeed the model's performance is inherently limited by the information in the low frequency component, and significant performance improvements should be archived by uncorrelated information from e.g. other statistics or expert judgements, as already discussed in section 6.3.

Despite strong correlations the bookmakers disagree remarkably more on the away selection than the other two selections, cf. figure 3.8. In addition the winning chances of strong away teams are generally underrated, cf. figures 5.11c and 5.11f, as a consequence of conservative odds setting. Presumably this is an expression of that the winning chances of away teams are generally difficult to estimate. Intuitively this is reasonable, as the home team advantage contributes negatively to the away team's strength and positively to the home team's strength. Essentially the home team is biased to win, and the away team must predominately rely of internal factors, such as morale and fatigue, which may be harder to estimate by the bookmakers.

Evidence suggests that bookmakers apply risk management procedures *individually* on each match by balancing the books. This risk-minimizing approach leads to inefficiencies due to biases an other irregularities in the bet distributions. Although these biases are well-known effects in virtually all betting markets, bookmakers accept the inefficiencies as the risk-minimizing approach minimizes the potential liability. One could however, speculate if such risk management is appropriate in the long run. If bookmakers instead act as profit-maximizers and accept losses on some matches, the profit would probably exceed that of the risk-minimizing approach. This can be regarded as a more *global* risk management. Presumably bookmakers are reluctant to bind large means despite that profit is statistically guaranteed in the long run, as it would require significantly larger cash capacity to accommodate potential series of adverse outcomes. This may explain the existence of many odds inefficiencies.

The simulations demonstrate that the model performance is highly seasons dependent, especially the basic model and the ensemble model profit very differently on each season. The posterior restrictive model also indicates that the regions with most lucrative bets vary according to seasons, cf. figures 5.10e and

5.10f. Additionally the odds analysis indicates an overround reduction. Obviously there are many contributing factors to this non-stationary behaviour, including the following:

Firstly the football sport evolves. New training techniques, strategic formations, talent development programs, etc. are enforced, as well as revised financial legislations, e.g. financial fair-play.

Secondly, the bookmaker industry evolves. Evidence indicates that the competition in the odds market is ever intensifying, as the overround is consistently decreasing in both leagues, cf. figure 3.3. Additionally bookmakers have become increasingly more accurate at setting lucrative books, as inefficient odds are increasingly penalized due to the intensified market competition [21]. This shows that odds setting models are constantly revised.

Thirdly, season and match specific circumstances apply. Coincidences such as injuries, suspensions, rowdy spectators, and pure luck during a match may influence the outcome. Additionally reinforcing mechanisms during a season may apply. A team may be superior in a long wave of matches due to increasingly high morale, or contrary be inferior in a wave due to reducing morale. Thirdly the match may be affected by match fixing possibly leading to unanticipated match results. A further meta-data analysis of seasons specific factors could have been carried out to elaborate on this subject.

6.5 Future Work

An interesting subject of further investigation is the team bias, as discussed in section 3.5. The current analysis demonstrates the existence of a team bias on the significantly strong and weak teams in both leagues by considering all engaged matches by all teams. A more selective analysis, where only particularly strong and weak teams are compared to each other, could possibly elaborate the results and give more relevant residual measures, as the matches unaffected by the bias are sorted out.

The team residual measure could further be applied to formulate a proximity measure between teams in matches with strong favourites, which should be used as an additional input feature. Instead of an explicit proximity measure one could also consider applying a binary input feature stating the home and away teams, whereby the model implicitly should detect the bias. Although this approach has been briefly examined with no apparent improvements in forecast precision, a more thorough analysis could have been performed.

The employed odds dataset consists of statistics from two independent sources. As discussed in section 3.1 the times at which these have been registered differs from each match, and the temporal odds movements are only available in the most recent season 12/13 at `betexplorer.com`. A further analysis of the temporal odds movements and its impact on the model performance could be of interests to illuminate if the overround changes and if opening or closing odds are generally better.

The decision framework may be revised as it applies the critical simplification that bet sizes are equal on engaged selections. Ideally the bets should be weighted to the expected return on the individual selections, such that large bets are made on single selections with high expected returns, and small best are made on selections with low expected returns. This modification drastically complicates the decision procedure, possibly leading to an optimization problem with respect to the weight of the bets on single selections which is very similar to Markowitz portfolio optimization, cf. section 1.5.

The ensemble model demonstrates that the selection bias can be reasonably handled by applying a 'max-min' strategy of the posteriors, where the standardized expected return is maximized with respect to combinations of selections, where the minimum posteriors from the ensemble is used. The results yield a reasonable reduction in the selection bias, but also indicate that the model occasionally understates the odds, suggesting a reverse selection bias. A further analysis and theoretical study could have been made on the impact of the approach, e.g. how does the number of members in the committee (networks) affect the bias.

The posterior restrictive model demonstrates high profitability capability and may be refined to improve the performance. In its current state the restrictions are primitively formulated with only eight possible intervals. A further analysis of the impact on different interval granularities could have been performed to optimize the flexibility. Alternatively a continuous penalty function may be applied to smoothly promote certain subintervals of the probabilistic range $[0, 1]$. Evidently the model oscillates between different subintervals I_{k_i}, cf. figures 5.10e and 5.10f, indicating more than one strongly lucrative subinterval is present.

The unit-less currency in the simulations emphasizes the generalizability of the results, as the bets size may be arbitrarily scaled to represent realistic wages while preserving the rate of return. Another, yet more interesting, adjustment of the bet sizes could be archived by applying a common betting system, such as a Martingale system, cf. section 1.4, where the bet sizes are time dependent. Thus one could simulate a full season with an initially fixed amount of money in a real currency, and obtain more tangible results in terms of profit.

Chapter 7

Conclusion

Bookmakers can be regarded as risk-adverse market providers, whose main objective is to obtain balanced books, whereby the profit per event ideally is indifferent to the outcome. Consequently, the odds must be adjusted according to the public opinion, which may lead to market inefficient odds. These adjustments reflect cognitive biases in common gambling behaviour. A special case is the 'favourite/long-shot' bias, whereby odds on strong favourites are overrated, and underrated on weak teams to compensate for a general tendency to bet on the risky weak team. Among the two considered leagues, the English Premier League and the Spanish La Liga, the bias is particularly strong in the La Liga, suggesting a more irrational betting behaviour in this league. A special variant of the bias is found on the individual teams, where strong teams, such as Real Madrid and Manchester United, are generally underrated. Additionally it has been demonstrated that bookmakers are biased to favour the home team, due to the home ground advantage. The data also indicates that odds setting techniques are temporarily non-stationary, as the overround – a measure of the bookmakers' profit safety margin – is consistently decreasing. This is most likely a consequence of increased market competition.

The proposed betting model uses the odds from a specified number of bookmakers as input features to a neural network and applies the emitted class posteriors in a decision framework. In this framework the standardized expected return per match is maximized by engaging an appropriate combination of outcome

selections. It has been demonstrated that the model is well-calibrated in both leagues, and that the accuracy is indifferent to whether 5, 9 or 37 bookmakers' odds are used as inputs. This indicates a generally low data complexity, and suggests that the variations between books at a given match basically are implications of odds balancing, which can be regarded as white noise.

Two extensions of the model have been proposed. Evidently the model is biased to engage selections where the posteriors are overestimated, and accordingly an ensemble method has been proposed, whereby the minimum posteriors from an ensemble of neural networks is applied in the decision framework, which reasonably efficiently removes the bias. Both the basic model and ensemble model are strongly season dependent in terms of profitability and practical application is questionable. The second extension relies on a restriction of the acceptable size of the class posteriors, yielding the generally highest rate of returns in both leagues. The model generally restricts the posteriors to $[0.4, 0.5]$ on recent seasons in both leagues. By nature people are poor probability estimators of intermediate probabilities and so bets on selections in this probability region are possibly strongly irregular, leading to significant odds inefficiencies. The model profits significantly more on the La Liga with a total profit/stake ratio of 16% across seasons 09/10-12/13, and is at least on par with former betting models in season 10/11 in the Premier League and seasons 09/10-10/11 in the La Liga.

Appendix A

Concepts and Operational Procedures in Odds Setting

A.1 Uniqueness of $\boldsymbol{b}$ in $\boldsymbol{\pi} = \Sigma\boldsymbol{b}$ in a balanced bet

Proof. Let σ_1, σ_2 and σ_3 denote the odds on home wins (result 1), draw (result 2) and away wins (result 3), respectively, for a given bookmaker X. Further let $B = b_1 + b_2 + b_3$ denote the total amount betted on X, with b_i, $i = 1, 2, 3$ being the bet on the i'th result. Assume that the overround $\eta = \sum_{i=1}^{3} \sigma_i^{-1} - 1 > 0$, with $\sigma_i > 1$, $i = 1, 2, 3$.

In order to make the bet balanced, the profit must be is fixed, regardless of the result. Denoting the fixed profit by π, the profit in case of result 1 can be stated as

$$\pi = b_2 + b_3 - b_1(\sigma_1 - 1) \tag{A.1}$$

Similar expressions can be made for results 2 and 3, yielding the following set of equations

$$\begin{bmatrix} \pi \\ \pi \\ \pi \end{bmatrix} = \begin{bmatrix} 1 - \sigma_1 & 1 & 1 \\ 1 & 1 - \sigma_2 & 1 \\ 1 & 1 & 1 - \sigma_3 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{A.2}$$

or written compactly,

$$\boldsymbol{\pi} = \Sigma \boldsymbol{b} \tag{A.3}$$

where $\boldsymbol{\pi} = [\pi, \pi, \pi]^T$, $\Sigma \in \mathbb{R}^{3\times 3}$ and $\boldsymbol{b} = [b_1, b_2, b_3]^T$. The determinant of the system matrix is given by

$$\det(\Sigma) = \sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_1\sigma_3 - \sigma_1\sigma_2\sigma_3 \tag{A.4}$$

By assumption $\eta > 0$. Hence

$$\eta = \frac{1}{\sigma_1} + \frac{1}{\sigma_2} + \frac{1}{\sigma_3} - 1 \tag{A.5}$$

$$= \frac{\sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_1\sigma_3 - \sigma_1\sigma_2\sigma_3}{\sigma_1\sigma_2\sigma_3} > 0 \tag{A.6}$$

By assumption $\sigma_i > 1$, $i = 1, 2, 3$ implying that $\sigma_1\sigma_2\sigma_3 > 0$ and therefore Eq. (A.6) can be reduced to

$$\sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_1\sigma_3 - \sigma_1\sigma_2\sigma_3 > 0 \tag{A.7}$$

Hence $\det(\Sigma) > 0$, provided that $\eta > 0$, and there exists a unique solution $\boldsymbol{b}$ to Eq. (A.3), ensuring that the bet is balanced. □

A.2 Relation between η, π and B in a balanced bet

Proof. The unique solution to Eq. (A.3), provided that $\eta > 0$, is given by

$$b_1 = \frac{\pi\sigma_2\sigma_3}{\sigma_2\sigma_3 + \sigma_1\sigma_3 + \sigma_1\sigma_2 - \sigma_1\sigma_2\sigma_3} \tag{A.8}$$

$$b_2 = \frac{\pi\sigma_1\sigma_3}{\sigma_2\sigma_3 + \sigma_1\sigma_3 + \sigma_1\sigma_2 - \sigma_1\sigma_2\sigma_3} \tag{A.9}$$

$$b_3 = \frac{\pi\sigma_1\sigma_2}{\sigma_2\sigma_3 + \sigma_1\sigma_3 + \sigma_1\sigma_2 - \sigma_1\sigma_2\sigma_3} \tag{A.10}$$

Since

$$\eta = \frac{1}{\sigma_1} + \frac{1}{\sigma_2} + \frac{1}{\sigma_3} - 1 \tag{A.11}$$

$$= \frac{\sigma_1\sigma_1 + \sigma_2\sigma_3 + \sigma_1\sigma_3 - \sigma_1\sigma_2\sigma_3}{\sigma_1\sigma_2\sigma_3} \quad \Leftrightarrow \tag{A.12}$$

$$\eta\sigma_1\sigma_2\sigma_3 = \sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_1\sigma_3 - \sigma_1\sigma_2\sigma_3 \tag{A.13}$$

Eq. (A.10) reduces to

$$b_1 = \frac{\pi\sigma_2\sigma_3}{\eta\sigma_1\sigma_2\sigma_3} = \frac{\pi}{\eta\sigma_1} \tag{A.14}$$

$$b_2 = \frac{\pi\sigma_1\sigma_3}{\eta\sigma_1\sigma_2\sigma_3} = \frac{\pi}{\eta\sigma_2} \tag{A.15}$$

$$b_3 = \frac{\pi\sigma_1\sigma_2}{\eta\sigma_1\sigma_2\sigma_3} = \frac{\pi}{\eta\sigma_3} \tag{A.16}$$

This gives

$$B = b_1 + b_2 + b_3 \tag{A.17}$$

$$= \frac{\pi}{\eta}\left(\frac{1}{\sigma_1} + \frac{1}{\sigma_2} + \frac{1}{\sigma_3}\right) \tag{A.18}$$

$$= \frac{\pi(\eta+1)}{\eta} \quad \Leftrightarrow \tag{A.19}$$

$$\pi = B\frac{\eta}{\eta+1} \tag{A.20}$$

This finishes the proof. □

Appendix B

Betting Strategies

B.1 Deriving expected gain and variance of gain

Let $i = 1, 2, 3$ denote the selections on a given match with home win ($i = 1$), draw ($i = 2$), and away win ($i = 3$). Assuming that the matches are independent, the expected gain $\mathrm{E}[\pi]$ and variance on the gain $\mathrm{Var}[\pi]$ can be derived individually on each match.

Consider now a single match. Let o_i, $i = 1, 2, 3$ denote the odds on selection i, and let $b \in \mathbb{R}^3$ denote a decision vector, where the ith element b_i, indicates how much to bet on selection i. Further, let $c \in \mathbb{R}^3$ denote a binary vector with the actual outcome of the match, where $c_i = 1$ indicates outcome i and $\sum_{i=1}^{3} c_i = 1$.

If no bets are made then obviously $E[\pi] = 0$. If *one* bet is made on selection $j \in \{1, 2, 3\}$, the gain π is

$$\pi = b_j (o_j c_j - 1) \tag{B.1}$$

and the expected gain is

$$\begin{aligned} E[\pi] &= E[b_j (o_j c_j - 1)] && \text{(B.2)} \\ &= b_j (o_j E[c_j] - 1) && \text{(B.3)} \\ &= b_j (o_j p_j - 1) && \text{(B.4)} \end{aligned}$$

where p_j denotes the probability of selection j as the outcome. Since

$$\begin{aligned} E\left[\pi^2\right] &= \mathrm{E}\left[b_j^2(o_j c_j - 1)^2\right] && \text{(B.5)}\\ &= \mathrm{E}\left[b_j^2 o_j^2 c_j^2 - 2b_j^2 o_j c_j^2 + b_j^2\right] && \text{(B.6)}\\ &= \mathrm{E}\left[b_j^2 o_j^2 c_j - 2b_j^2 o_j c_j + b_j^2\right] && \text{(B.7)}\\ &= b_j^2 o_j^2 \mathrm{E}\left[c_j\right] - 2b_j^2 o_j \mathrm{E}\left[c_j\right] + b_j^2 && \text{(B.8)}\\ &= b_j^2 o_j^2 p_j - 2b_j^2 o_j p_j + b_j^2, && \text{(B.9)} \end{aligned}$$

the variance can be expressed as

$$\begin{aligned} \mathrm{Var}\left[\pi\right] &= \mathrm{E}\left[\pi^2\right] - \mathrm{E}^2\left[\pi\right] && \text{(B.10)}\\ &= b_j^2 o_j^2 p_j - 2b_j^2 o_j p_j + b_j^2 - b_j^2\left(o_j p_j - 1\right)^2 && \text{(B.11)}\\ &= b_j^2 o_j^2 p_j - 2b_j^2 o_j p_j + b_j^2 - \left(b_j^2 o_j^2 p_j^2 - 2b_j^2 o_j p_j + b_j^2\right) && \text{(B.12)}\\ &= b_j^2 o_j^2 p_j (1 - p_j) && \text{(B.13)} \end{aligned}$$

Assume now that *two* bets are made on distinct selections j and k, $j, k \in \{1, 2, 3\}$. For simplicity also assume that the same amount is waged on each selection, i.e. $b = b_j = b_k$. The gain is then

$$\pi = \pi_j + \pi_k = b\left(o_j c_j - 1\right) + b\left(o_k c_k - 1\right) = b\left(o_j c_j + o_k c_k - 2\right) \qquad \text{(B.14)}$$

and consequently the expected gain, cf. Eq. (B.4) is

$$\mathrm{E}\left[\pi\right] = \mathrm{E}\left[\pi_j\right] + \mathrm{E}\left[\pi_k\right] = b\left(o_j p_j + o_k p_k - 2\right) \qquad \text{(B.15)}$$

Exploiting that $c_j c_k = 0$, $j \neq k$ and $c_i^2 = c_i$, $i = 1, 2, 3$ yields

$$\begin{aligned} \mathrm{E}\left[\pi^2\right] &= \mathrm{E}\left[b^2\left(o_j c_j + o_k c_k - 2\right)^2\right] && \text{(B.16)}\\ &= \mathrm{E}\left[b^2 o_j^2 c_j^2 - 4b^2 o_j c_j + 2b^2 o_j c_j o_k c_k + 4b^2 - 4b^2 o_k c_k + b^2 o_k^2 c_k^2\right] && \text{(B.17)}\\ &= b^2 o_j^2 \mathrm{E}\left[c_j\right] - 4b^2 o_j \mathrm{E}\left[c_j\right] + 4b^2 - 4b^2 o_k \mathrm{E}\left[c_k\right] + b^2 o_k^2 \mathrm{E}\left[c_k\right] && \text{(B.18)}\\ &= b^2 o_j^2 p_j - 4b^2 o_j p_j + 4b^2 - 4b^2 o_k p_k + b^2 o_k^2 p_k, && \text{(B.19)} \end{aligned}$$

and since

$$\mathrm{E}^2\left[\pi\right] = b^2\left(o_j p_j + o_k p_k - 2\right)^2, \qquad \text{(B.20)}$$

the variance can be expressed as

$$\begin{aligned} \mathrm{Var}\left[\pi\right] &= \mathrm{E}\left[\pi^2\right] - \mathrm{E}^2\left[\pi\right] && \text{(B.21)}\\ &= b^2 o_j^2 p_j - b^2 o_j^2 p_j^2 + b^2 o_k^2 p_k - b^2 o_k^2 p_k^2 - 2b^2 o_j o_k p_j p_k && \text{(B.22)}\\ &= \mathrm{Var}\left[\pi_j\right] + \mathrm{Var}\left[\pi_k\right] - 2b^2 o_j o_k p_j p_k && \text{(B.23)} \end{aligned}$$

This finishes the derivation.

Appendix C

Model Definition

C.1 Derivation of cost function gradient

Following the notation in section 4.1, application of the chain rule yields

$$\frac{\delta E_n}{\delta w_{ji}} = \frac{\delta E_n}{\delta a_j} \frac{\delta a_j}{\delta w_{ji}} \tag{C.1}$$

For convenience, denote

$$\delta_j^h = \frac{\delta E_n}{\delta a_j^h} \tag{C.2}$$

$$\delta_j^o = \frac{\delta E_n}{\delta a_j^o} \tag{C.3}$$

First, consider $\frac{\delta E_n}{\delta w_{ji}^o}$. By Eq. (4.3) it is seen that

$$\frac{\delta a_j^o}{\delta w_{ji}^o} = z_i \tag{C.4}$$

Repeated use of the chain rule yields

$$\delta_j^o = \sum_{k=1}^{N_y} \frac{\delta E_n}{\delta y_k} \frac{\delta y_k}{\delta a_j^o} \tag{C.5}$$

It is noted that the partial derivatives of the softmax function, y_k, cf. Eq. (4.2) are given by [22, p. 209],

$$\frac{\delta y_k}{\delta a_j^o} = y_k \left(I_{kj} - y_j \right) \tag{C.6}$$

where I_{kj} are the elements of the identity matrix, and

$$\frac{\delta E_n}{\delta y_k} = \frac{\delta}{\delta y_k} \left(- \sum_{k'=1}^{C} t_{k'} \ln y_{k'} \right) = - \frac{t_k}{y_k} \tag{C.7}$$

Combining Eqs. (C.6) and (C.7) gives

$$\delta_j^o = - \sum_{k=1}^{N_y} \frac{t_k}{y_k} y_k \left(I_{kj} - y_j \right) = - \left(t_j - y_j \sum_{k=1}^{N_y} t_k \right) = y_j - t_j \tag{C.8}$$

Hence, using Eqs. (C.4) and (C.8),

$$\frac{\delta E_n}{\delta w_{ji}^o} = (y_j - t_j) z_i \tag{C.9}$$

Now consider $\frac{\delta E_n}{\delta w_{ji}^h}$. By Eq. (4.1) it follows that

$$\frac{\delta a_j^h}{\delta w_{ji}^h} = x_i \tag{C.10}$$

Repeated use of the chain rule yields

$$\delta_j^h = \sum_{k=1}^{N_y} \frac{\delta E_n}{\delta a_k^o} \frac{\delta a_k^o}{\delta a_j^h} \tag{C.11}$$

By combining Eqs. (4.1) and (4.3) it is seen that

$$a_k^o = \sum_{i=0}^{N_z} w_{ki}^o h(a_i^h) \quad \Rightarrow \quad \frac{\delta a_k^o}{\delta a_j^h} = h'(a_j^h) w_{kj}^o \tag{C.12}$$

Using Eqs. (C.3), (C.11), and (C.12) yields

$$\delta_j^h = h'(a_j^h) \sum_{k=0}^{N_y} \delta_k^o w_{kj}^o \tag{C.13}$$

whereby it is finally deduced from Eqs. (C.10) and (C.13) that

$$\frac{\delta E_n}{\delta w_{ji}^h} = \left(h'(a_j^h) \sum_{k=0}^{N_y} w_{kj}^o \delta_k^o \right) x_i \tag{C.14}$$

This finishes the derivation. [22, pp. 241-245],[44, pp. 39-42]

Bibliography

[1] The history of sports betting. www.onlinecasinoadvice.com/sportsbetting/history/. Accessed: 20/03/2013.

[2] Roger Munting. *An Economic and Social History of Gambling in Britain and the USA*. Manchester University Press, 1996. ISBN 0719044499.

[3] The history of bookmakers. http://ezinearticles.com/?The-History-of-Bookmakers&id=2177406. Accessed: 20/03/2013.

[4] The race track gangs. www.epsomandewellhistoryexplorer.org.uk/RaceTrackGangs.html. Accessed: 20/03/2013.

[5] The "big three" – william hill, ladbrokes and coral. www.reliablebookies.com/bookmakers/the-big-three-william-hill-ladbrokes-and-coral/. Accessed: 20/03/2013.

[6] Addiction soars as online gambling hits £2bn mark. www.independent.co.uk/news/uk/home-news/addiction-soars-as-online-gambling-hits-2bn-mark-8468376.html. Accessed: 20/03/2013.

[7] Bookmaker history. www.britishbookmakers.co.uk/bookmakers/bookmaker-history.htm. Accessed: 20/03/2013.

[8] Gambling online illegal in europe? www.latestcasinobonuses.com/is-gambling-online-illegal-in-europe.html. Accessed: 20/03/2013.

[9] Danske licens spil. danskespil.dk/om/koncernen/dli/. Accessed: 20/03/2013.

[10] Skat: Slut med spil-monopol. politiken.dk/indland/ECE1398473/skat-slut-med-spil-monopol/. Accessed: 20/03/2013.

[11] Usa gambling laws. online-gambling-legal.info/usa-gambling-laws.php. Accessed: 20/03/2013.

[12] Legality of gambling in various countries. http://online-gambling-legal.info/. Accessed: 20/03/2013.

[13] Betting exchange definition. www.soccernews.com/soccer-betting/betting-glossary/betting-exchange/. Accessed: 20/03/2013.

[14] Betting exchanges. www.online-betting.me.uk/betting-exchanges.html. Accessed: 20/03/2013.

[15] J. Moroney. *Facts from figures.* Pelican books ; A236. Penguin Books, 1951. LCCN 52003716.

[16] M.J. Maher. Modelling association football scores. *Statistica Neerlandica, Nr. 36*, pages 109–118, 1982.

[17] R. Pollard C. Reep and B. Benjamin. Skill and chance in ball games. *Journal of the Royal Statistical Society: Series A (General), Vol. 134, No. 4*, pages 623–629, 1971.

[18] I. D. Hill. Association football and statistical inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol. 23, No. 2*, pages 203–208, 1974.

[19] M. Dixon and S. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol. 46, No. 2*, pages 265–280, 1997.

[20] T. Kuypers. Information and efficiency: an empirical study of a fixed odds betting market. *Applied Economics, Vol. 32, No. 32*, pages 1353–1363, 2000.

[21] J. Goddard D. Forrest and R. Simmons. Odds-setters as forecasters: The case of english football. *International Journal of Forecasting, Vol. 21, No. 3*, pages 551–564, 2005.

[22] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006. ISBN 0387310738.

[23] N. Fenton A. Joseph and M. Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems, Vol. 19, No. 7*, page 544–553, 2005.

[24] N. E. Fenton A. C. Constantinou and M. Neil. A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems, Vol. 36*, pages 322–339, 2012.

[25] Hannes Anderson. A machine learning approach to sports betting, 2011. DTU Course 02459.

[26] Anti-martingale system. `www.investopedia.com/terms/a/antimartingale.asp`. Accessed: 14/05/2013.

[27] Martingales. `www.math.nyu.edu/faculty/varadhan/course/PROB.ch5.pdf`. Accessed: 14/05/2013.

[28] David Williams. *Probability with Martingales*. Cambridge University Press, 1991. ISBN 0521406056.

[29] Reverse martinale system. `www.reversemartingale.com`. Accessed: 14/05/2013.

[30] Gambler's fallacy. `www.fallacyfiles.org/gamblers.html`. Accessed: 14/05/2013.

[31] Steven Roman. *Introduction to the Mathematics of Finance: From Risk Management to Options Pricing*. Springer-Verlag New York, LLC, 1st edition, 2004. ISBN 9780387213644. 369 pp.

[32] Methods of representing odds. `betstarter.com/sportsbetting/OddsTypes.asp`. Accessed: 17/06/2013.

[33] Popular football bets. `www.betting-directory.com/football/popular-football-bets.php`. Accessed: 20/03/2013.

[34] Fixed-odds betting. `www.sportinglybetter.com/betting-terms/each-way-betting-explained/`. Accessed: 20/03/2013.

[35] Bet types available. `support.skybet.com/app/answers/detail/a_id/11/~/bet-types-available`. Accessed: 17/06/2013.

[36] The art of bookmaking. `betting.betfair.com/education/-generic-betting-principles/the-art-of-bookmaking-1-110111.html`. Accessed: 20/03/2013.

[37] Bookmaker. `http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Bookmaker.html`. Accessed: 20/03/2013.

[38] List of premier league winners. `www.soccer-blogger.com/2011/06/27/list-of-premier-league-winners-1992-till-date/`. Accessed: 26/06/2013.

[39] List of spain la liga champions. english.ahram.org.eg/NewsContent/6/55/71229/Sports/World/List-of-Spain-La-Liga-champions.aspx. Accessed: 26/06/2013.

[40] Bookmakers by year established. www.top100bookmakers.com/established/. Accessed: 06/06/2013.

[41] Backpropation learning: Exercise 5, 2012. Course material from DTU course 02457.

[42] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review, Vol. 78, No. 1*, pages 1–3, 1950.

[43] Steinbach M. Kumar V. Tan, P. N. *Introduction to Data Mining*. Pearson Education, Inc., 2006. ISBN 0321420527.

[44] Lecture 6: Perceptrons for signal detection, 2012. Course material from DTU course 02457.