SPORT PREDICTION MARKET MODELING WITH PORTFOLIO OPTIMIZATION

ONDŘEJ HUBÁČEK

Doctoral Thesis

Ph.D. Programme: Electrical Engineering and Information Technology
Branch of Study: Artificial Intelligence and Biocybernetics

Department of Computer Science
Faculty of Electrical Engineering
Czech Technical University in Prague

Supervisor: prof. Ing. Filip Železný, Ph.D.
Supervisor-specialist: Ing. Gustav Šír, Ph.D.

2024

Ondřej Hubáček: *Sport prediction market modeling with portfolio optimization,*
a thesis submitted for the degree of Doctor of Philosophy (Ph.D.),
Study Programme No. P2612-Electrotechnics and Informatics,
Branch No. 3902V035 - Artificial Intelligence and Biocybernetics.

SUPERVISORS:
prof. Ing. Filip Železný, Ph.D.
Ing. Gustav Šír, Ph.D.

LOCATION:
Prague, Czech Republic

DECLARATION

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published work of others has been acknowledged in the text, and a list of references is given.

Prague, Czech Republic, February 2024

Ondřej Hubáček

ABSTRACT

Over the past decade, the surge in the use of machine learning models has revolutionized various domains, witnessing notable advancements in areas such as image recognition, reinforcement learning in games, machine translation, and language generation. This thesis extends the application of machine learning techniques to forecasting, a domain that remains relatively unexplored. Specifically, our focus is on predicting the outcomes of sports events and leveraging these predictions in trading on prediction markets.

The initial phase of our investigation involves a comprehensive review of the state-of-the-art in score-based modeling. Despite the existence of seemingly diverse methods, a quantitative comparison on large-scale data is lacking. To address this gap, we re-implement and benchmark nine existing models using the largest publicly available dataset. Our evaluation framework ensures a fair comparison, revealing that the predictive performance of these models is remarkably similar. Further analysis of the predictions highlights that this similarity is predominantly due to inherent similarities in the prediction outputs.

After establishing a baseline for our endeavor we designed and implemented our own models, testing two distinct approaches. One approach relies on carefully engineered score-derived features, while the other capitalizes on the relational structure of the data. The feature-based classifier outperforms state-of-the-art models by a significant margin across all examined metrics. We showcase the model's adaptability by seamlessly integrating outputs from a simpler model as additional inputs to the classifier, achieving notable improvements through feature engineering. Despite these advancements, our model lags significantly behind bookmakers' predictions, suggesting the need for more complex models or a reevaluation of the overarching goal of achieving universally more precise predictions than the market.

Turning our attention to trading our predictions on the markets, we design a neural model tailored for the NBA competition, utilizing detailed player-level data from each game. Departing from the traditional accuracy-based approach to forecasting, we introduce the concept of decorrelation as a method for profiting on the markets using a model with inferior performance by conventional metrics. Additionally, we formalize the often-neglected concept of market taker's advantage. To validate these concepts, we subject them to testing through simulations and real-world data. The results demonstrate that the decorrelation is an effective way to achieve profits.

Keywords: machine learning, prediction markets, forecasting in sports, decision making, gradient boosted trees, neural networks

ABSTRAKT

Během poslední dekády došlo k nárůstu využívání modelů strojového učení, které přinesly revoluci v různych doménách, s významnými pokroky v oblastech jako rozpoznávání obrazu, posilované učení ve hrách, strojový překlad či generování jazyka. Tato disertační práce rozšiřuje aplikaci technik strojového učení na predikování, oblast, která zůstává relativně neprozkoumána. Konkrétně se zaměřujeme na predikci výsledků sportovních událostí a využívání těchto predikcí k obchodování na predikčních trzích.

V počáteční fázi našeho výzkumu se zabýváme komplexním přehledem aktuálních poznatků ohledně modelování na základě výsledných skóre utkání. Navzdory existenci zdánlivě rozličných metod chybí jejich kvantitativní srovnání na velkém vzorku dat. Abychom toto srovnání doplnili, reimplementujeme a porovnáváme devět existujících modelů na největší veřejně dostupné datové sadě. Náš vyhodnocovací framework zajišťuje spravedlivé porovnání a odhalujeme, že prediktivní kapacita zkoumaných modelů je velmi podobná. Další analýza predikcí modelů naznačuje, že tato podobnost je převážně způsobena podobnostmi samotných predikcí těchto modelů.

Po kvantitativním vyhodnocení stávajících modelů navrhujeme a implementujeme vlastní modely, založené na dvou odlišných přístupech. Jeden se spoléhá na pečlivě navržené příznaky odvozené ze skóre, zatímco druhý využívá relační strukturu dat. Klasifikátor založený na odvozených příznacích překonává stávající modely se značným odstupem ve všech zkoumaných metrikách. Integrací výstupu jednodušších modelů demonstrujeme flexibilitu klasifikátoru, který pomocí dalších příznaků dosahuje výrazného zlepšení. Navzdory těmto výsledkům náš model výrazně zaostává za predikcemi sázkových kanceláří, což naznačuje, že se neobejdeme bez komplexnejších modelů nebo přehodnocení celkového cíle dosažení univerzálně přesnějších předpovědí, než kterými disponuje trh.

V další části přesouváme naši pozornost k obchodování našich predikcí na trzích. Navrhujeme neuronovou síť přizpůsobenou soutěži NBA, využívající detailní data o hráčích z každého zápasu. Odchylujeme se od konvenčního přístupu zaměřeného výlučně na přesnost predikcí a zavádíme koncept dekorelace jako metodu pro porážení trhů. Dále formálně definujeme často opomíjený koncept výhody tzv. market-takera. Tyto koncepty testujeme pomocí simulací a reálných dat. Výsledky jednoznačně dokazují účinnost navržených metod.

Klíčová slova: strojové učení, prediktivní trhy, predikování ve sportu, rozhodovací proces, gradientní rozhodovací stromy, neuronové sítě

PUBLICATIONS

Some contents of this thesis have appeared previously in the following publications :

Hubáček, Ondřej and Gustav Šourek. "Beating the market with a bad predictive model." In: *International Journal of Forecasting* 39.2 (2023), pp. 691–719

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. "Forty years of score-based soccer match outcome prediction: an experimental review." In: *IMA Journal of Management Mathematics* 33.1 (2021), pp. 1–18

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. "Exploiting sports-betting market using machine learning." In: *International Journal of Forecasting* 35.2 (2019), pp. 783–796

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. "Learning to predict soccer results from relational data with gradient boosted trees." In: *Machine Learning* 108.1 (2019), pp. 29–47

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. "Lifted Relational Team Embeddings for Predictive Sport Analytics." In: *CEUR Workshop Proceedings*. Vol. 2206. CEUR-WS, 2018, pp. 84–91

You are a very fine person, Mr Baggins,
and I am very fond of you;
but you are only quite a little fellow
in a wide world after all!

— Gandalf

ACKNOWLEDGMENTS

First of all, I would like to to express my sincere gratitude to my supervisor specialist Gustav Šír, who provided me a guidance through my doctoral studies. Secondly, my thanks goes to my supervisor Filip Železný for his continuous support and unwavering patience.

From elementary school to university, I had the privilege of learning from dedicated educators who helped me develop my interest for various subjects. I am particularly grateful to Petr Špína, Luděk Absolon, and Petr Drahotský.

Special appreciation goes to my colleagues from Ematiq, who helped me rediscover my passion for sports trading.

My warmest thanks extend to my family for their support through my whole studies. Finally, my deepest gratitude goes to my beloved wife, Veronika, who stood by my side through this entire journey.

CONTENTS

III Proposed Machine Learning Models

IV Beating the Bookies

V Conclusions

VI Appendix

Part I

INTRODUCTION

1 INTRODUCTION

As artificial intelligence continues its steady integration into diverse domains, our ambition is to leverage the capabilities of machine learning models for forecasting, with a specific emphasis on predicting the outcomes of sports events and trading these predictions on the markets.

The history of sports betting can be traced back to Ancient Greece and still attracts interest, as seen in the ongoing activity on the markets. We can imply that forecasting in sports is not a solved problem otherwise there would be no incentives to trade. Unlike ancient Greeks, we can rely on the computational power of modern computers and various historical data for our data-driven approach.

The thesis is organized as follows. We introduce the problem and underlying terms in Part i. Part ii presents an experimental review of state-of-the-art models for score-based match outcome modeling. In Part iii, we design and implement our own models and compare them to the state-of-the-art. Finally, in Part iv we focus on trading our predictions on the markets.

1.1 PROBLEM STATEMENT

Our research aims to utilize the potential of machine learning models for profit generation. To achieve this objective, we should identify suitable domains, collect historical data, design and develop machine learning models, establish an evaluation framework, and explore effective strategies for trading the predictions on the markets.

The overarching goal of our study is twofold. Firstly, we seek to assess the predictive accuracy of these machine learning models by benchmarking them against state-of-the-art standards. Secondly, our research extends beyond predictive accuracy to evaluate the practical profitability of these models. By comparing their performance against market data, we investigate how well the predictions can be translated into financial gains in a real-world trading scenario.

1.2 CONTRIBUTIONS

We have experimentally evaluated a class of score-based models used for forecasting the winner of a soccer match (Part ii). While there is a large body of literature centered around this topic, surprisingly, a comparative study has not been done. The review was conducted on the largest publicly available dataset and can serve as a benchmark for future endeavors in this domain.

In Part iii, we implemented a model improving upon the state-of-the-art performance. Despite the closely matched performance of the reviewed models, our classifier demonstrated a substantial improvement (Section 12.7). Additionally, we challenged the use of the Ranked Probability Score as the sole metric for evaluating predictions in games with a low prior probability of draws (Section 13).

In Part iv, we introduced the concept of decorrelation. Our findings revealed that profits on prediction markets can be attained through means beyond merely possessing a more accurate model (Section 17). A market taker can capitalize on his/her advantage (Section 16.3) when he/she finds a compromise between accurate and dissimilar predictions from the market. Moreover, we demonstrated that such predictions can be obtained by setting the right incentives during the model fitting (Section 18.3). Last but not least, we showed that the Modern Portfolio Theory framework works well with the introduced concepts, leading to a substantial improvement over the baseline uniform investment strategy (Section 19.4).

2

BACKGROUND

This background section introduces the problem setup (Section 2.1), and basic concepts from betting markets, machine learning, and portfolio optimization.

2.1 PROBLEM SETUP

During a sports event, various markets are traded, each representing a different aspect of the game. For example, in a soccer match, markets such as the *Moneyline* and *Totals* are commonly traded. The Moneyline market represents the probability of which team will win the match, while the Totals market trades the probability of a certain number of goals being scored. Within each market, there are selections that traders can choose from, such as "Home" or "Away" for the Moneyline market or "Over 2.5 goals" or "Under 2.5 goals"for the Totals market. These selections represent the tradeable assets.

OPPORTUNITY In a market, each tradeable selection $s \in \mathcal{S}$ corresponds to an opportunity to back (buy, α) and lay (sell, β) the selection, respectively. We will further refer to all existing possibilities to trade a certain selection s jointly as *opportunities* $\omega_s^\zeta \in \Omega$. [1] Nevertheless, where necessary, we will distinguish the side of the market $\zeta \in \{\alpha, \beta\}$ an opportunity ω_i^ζ is to be traded at.

PRICE Each opportunity ω_i is then associated with a certain price. The price reflects the market's perceived probability of an outcome to occur. It is also commonly referred to as "odds" that determine the potential payouts received from a wager.

FAIR PRICE We assume that each opportunity ω_i, which corresponds to a selection being traded, has an underlying fundamental fair price r_i. This fair price can be expressed as a real number from the interval $[0, 1]$ and can be compared to some price estimate in the same unit of measurement (Section 2.5). The fair price (odds) reflects

[1] i.e. two opportunities ω_i and ω_j over the same selection at two different times will be treated the same as two opportunities ω_i and ω_j over two different selections. We note that there typically are dependencies between the opportunities that could theoretically be used to further improve the underlying techniques, however we do not exploit these in this thesis.

symbol	**description**
$s_j \in \mathcal{S}$	selections being traded on the market
$\{\alpha, \beta\};\ \zeta \in \mathrm{Z}$	market sides, back and lay
$\omega^{\zeta}_{s_j};\ \omega_i \in \Omega$	opportunities to trade selection
$\mathcal{D}_i \subset \mathcal{D}^*$	relevant data available for opportunity ω_i
$\mathsf{r} : \mathcal{D}^*_i \rightarrow [0, 1]$	fair price r of a selection
$\mathsf{m} : \mathcal{D}^b_i \rightarrow [0, 1]$	market maker m, or simply "the market"
$\mathsf{t} : \mathcal{D}^m_i \rightarrow [0, 1]$	market taker t
P_{Ω}	distribution of market opportunities
$R : \omega_i \mapsto r_i$	r.v. denoting selection fair price
$M : \omega_i \mapsto m_i$	r.v. denoting market maker's pricing
$T : \omega_i \mapsto t_i$	r.v. denoting market taker's pricing
$\mathrm{P}(R, M, T)$	distribution of the price estimates (values)
$\theta \in \Theta$	model parameters
$\tilde{\mathsf{e}} : \mathcal{D}_i \rightarrow \tilde{\mathrm{E}}_\mathrm{i}$	posterior distribution $\tilde{\mathrm{E}}$ estimator $\tilde{\mathsf{e}}$ (e.g. $\tilde{\mathsf{t}}$)
$\rho_i \in \mathbb{R}$	trading returns
$\mathsf{W} \in \mathbb{R}_+$	wealth of the market taker
$\boldsymbol{f} \in \mathbb{R}^n$	vector of the wealth allocations

Table 2.1: Overview of the used notation.

the true (inverse) probability of the associated outcome occurring. Note that the fair price of a selection s is always the same for both sides (α, β) of the market. While the concept of a fair price r_i can be seen as somewhat speculative, since it cannot be directly measured, we note that we merely require its theoretical existence for defining market efficiency (Section 2.2) and the related concepts (Section 2.6).

MAKERS AND TAKERS We commonly distinguish two types of the market participants as (i) market *makers* m, and (ii) market takers t. The market makers continuously quote both sides (α, β) of the market at certain price levels m_i resulting into trading opportunities $\omega_i \in \Omega$. The market takers are then *selecting* from the existing opportunities ω_i to issue specific back and lay orders. We generally consider the problem analysed in this paper as a two-player game between a market maker m and taker t, where each player possesses a certain pricing policy $\Omega \rightarrow [0, 1]$ over the market distribution of opportunities P_Ω given by the game (world) environment.

RESULTING When the market taker t accepts the market maker's m price m_{ii} their orders match. Besides the price, each order includes a backer's stake f_{max}. The matched amount f_i is the minimum f_{max} from the back and lay orders.[2] Once the event concludes, the selections are assigned either *win* or *loss* result. The backers of the winning selections receive a payoff $f_i \cdot \frac{1}{m_{ii}}$, leading to the profit of $f_i \cdot \frac{1}{m_{ii}} - f_i$ as they already paid f_i to the layer. The backers of the losing selections lose their stakes to the layers.[3]

BEATING THE MARKET We then aim to design a strategy for the role of the market *taker* to make positive profits against some particular market maker m. For example, such strategy would allow a trader to make consistent long-term profits while wagering against a particular bookmaker. Note that this is a zero-sum setting, where the profits of the market taker are at the direct expense of the market maker, as the fair price is objectively the same for all the participants. In this thesis, we then utilize the phrase "beating the market" to refer to profiting in this competitive setting.

2.2 MARKET EFFICIENCY

In a (strongly) efficient market, the current market price $m \in [0, 1]$ of a selection s reflects all existing information, making it impossible for any trader to make consistent profit by outsmarting the market[4] [37]. Particularly, the price m_i of each opportunity ω_i being traded needs to be an *unbiased* estimate (Section 2.5) of its fair price r_i [105]. Note that this does not imply the price m_i to be equal to the fair price r_i, but merely

[2] For simplicity, we further assume, that $f_i = f^t_{max}$

[3] We acknowledge that the there are markets, where the resulting is more complex, but we do not deal with those in this thesis.

[4] Note that this does not imply that one needs comparably more information to beat an inefficient market.

that the error of the estimation $m_i = \hat{r}_i$ is fully due to an irreducible variance which is *completely random* (Section 2.5) [113]. The inherent randomness of the error then ensures that no trader can consistently beat the market to secure systematic profits.

In real world liquid markets, profit-maximizing traders continuously search for misspriced oportunities to secure their profits, pushing the market price to quickly converge to the fair price in the process. Thanks to this self-regulating mechanism, the market inefficiencies tend to vanish quickly over time [38, 120]. We note that the idea of an efficient market is a rather theoretical one, since in a market that would become completely efficient, the traders would loose the incentive to search for inefficiencies, in turn making the market inefficient again. Any real world market can thus be hardly considered as perfectly efficient, nevertheless some measurable degree of efficiency, such as statistical unbiasedness of the mean prices (Section 16.4), is typically present in liquid markets [42]. In this work, we will further consider the most realistic setting of a partially (in-)efficient market where the market price m_i is a very good estimate of the fair price r_i, but not a perfect one.

2.3 MARKET MAKER'S ADVANTAGE

Market makers are essential to trading by providing constant liquidity to both sides (α, β) of the market, for which they are typically favored by the exchange operator in terms of fees and commissions. However, the position of a market maker is a difficult one, since he/she needs to constantly price the selections as accurately as possible.

To improve his/her position and secure profit, the market maker incorporates a so called *spread* ϵ into his/her estimates M, causing the offered price to back and lay the opportunity ω_i to differ some ϵ. The spread ϵ then works as a safety patch on the market maker's estimation error, preventing from one-sided exploitations by the market takers aiming at the discrepancy from the fair price. A safe strategy for the market maker is then to set his/her estimate and spread such that $\forall i : {m_i}^{\alpha} < r_i < {m_i}^{\beta}$, making it impossible for any trader to make any profit.

However, given that $m = \hat{r}$ is merely an estimate, it is still possible for the fair price r to fall outside the (m^{α}, m^{β}) region. [5] To further mitigate possible exploitation by the traders, the market maker can continuously adapt his/her estimate to the traders' behavior. That is he/she can responsively move his/her estimate m once the demand of one side of the market starts to prevail, indicating expected value perceived by the traders, stemming from a possibly erroneous price estimate m. In the ideal case where he/she is continuously able to maintain a perfectly balanced book, he/she is again guaranteed a profit of ϵ per pair of trades. Note that the market maker's profit in this case is independent of the fair price r.

[5] Naturally, this could be mitigated by increasing the ϵ, however a margin set too large will discourage market takers from trading, consequently removing the market maker's profit, too.

On the other hand, the market maker can theoretically digress from this purely reactive position to actively speculate against the takers' opinions and aim at a profit even higher than ϵ, at the cost of involving risk stemming from his/her, possibly erroneous, estimate m. This effectively allows the market maker to speculate on the fair price, which can lead to higher expected profits in settings where he/she possesses a superior price prediction model. [6] This is common, for instance, with bookmakers in the sport prediction markets [96]. Naturally, the market makers can also combine the aforementioned methods.

We note we often omit the spread ϵ from calculations and simulations further in this paper for simplicity of explanation[7], since it does not affect the main principles introduced in this thesis. The underlying assumption is that the spread constitutes an independent offset on the market prices, hence decreasing the resulting profits, but not interfering with the price estimation problem itself.

2.4 MARKET MODELING

Market modeling generally refers to the approach of fitting a statistical estimator e to the available market-related data $\mathcal{D}$ in order to capture its true distribution P. Possession of such a model then enables to answer all sorts of statistical queries, including the essential estimation of the fair prices R of market opportunities Ω.

The quality of the estimation of the fair prices then follows directly from two factors, (i) quality of the data $\mathcal{D}^t$ available to the trader t, which can possibly provide superior information which is not reflected in the market price, and (ii) quality of the model t used to fit the data, since different information w.r.t. r can possibly be derived with different methods, despite the same data source $\mathcal{D}^t$ given.

DATA The market-related data $\mathcal{D}$ may come from various external sources such as news and collected match statistics, as well as from the market itself (e.g. the traders' behavior). Recall that in an efficient market, all relevant data $\mathcal{D}^*$ must be used by the market maker m for pricing of each opportunity $\forall i : \mathsf{m}(\mathcal{D}_i^*) \mapsto m_i$, however, in practice it is more likely that $\mathcal{D}_i^m \subset \mathcal{D}_i^*$. The market takers can theoretically use the same data as the market maker $\mathcal{D}^t = \mathcal{D}^m$, although their sources are typically more limited (e.g. Section 19.2). However, they commonly strive to gain at least some information advantage by obtaining data which are not available to m, i.e. $\mathcal{D}^t \setminus \mathcal{D}^m \neq \emptyset$, and thus not reflected in the market price. Generally if $\mathcal{D}^t \subset \mathcal{D}^m$, such information advantage is completely missing, making it impossible to beat the market through superior price estimations[8], unless using a superior model.

[6] Since the market maker typically needs an in-depth knowledge of the market to operate, he/she can often reasonably expect to also possess price estimates superior to the average trader.

[7] except for the actual experiments with real data (Section 19).

[8] However, we note that it is still possible to make profits in such a scenario (Section 16.2).

MODELING The models used to fit the data are generally some θ-parameterized functions, the properties of which also influence the quality of the estimation. Ultimately, one would strive to model the whole joint distribution over the data $\mathcal{D}$, enabling to answer all possible probabilistic queries about the domain, consequently leading to truly optimal decision making. Nevertheless, the common investment strategies (Section 2.7) are typically based merely on estimates of returns from the opportunities ω_i at hand, which restricts the task to modeling of a conditional of the fair price $\mathrm{P}_\Omega(R|\mathcal{D}_i)$.

For instance, an estimate of the fair price r_i in the betting market can be derived from repeated observations of the outcomes of the associated stochastic events aggregated over a large enough sample from the market.

MISPRICING We assume a market that is not fully efficient (Section 2.2), i.e. there must be some mispriced opportunities which further need to be identifiable by some systematic means. The goal of the traders is then to identify such opportunities where $r_i \notin ({m_i}^\alpha, {m_i}^\beta)$ by comparing the market price m_i to their own estimate t_i. Should the market price m_i for an opportunity ω_i actually deviate from the fair price r_i in a non-random manner, such mispricing can be turned into positive returns.

2.5 PRICE ESTIMATORS

A price estimator e is generally a θ-parameterized function mapping some input data $\mathcal{D}_i$ associated with an opportunity ω_i onto a point price estimate $\hat{r}_i \in [0,1]$

$$\mathsf{e} : \mathcal{D}_i \mapsto \hat{r}_i \tag{2.1}$$

However, every prediction is associated with a certain level of uncertainty, which is either inherent or stemming from the missing information at the time of making ($\mathcal{D}_i^t \subset \mathcal{D}_i^*$). One might thus want to quantify the uncertainty by associating each possible estimate $\hat{r}_i$ for an opportunity ω_i with a probability, resulting into a posterior distribution estimation

$$\tilde{\mathsf{e}} : \mathcal{D}_i \mapsto \tilde{\mathrm{R}}_i \tag{2.2}$$

where $\tilde{\mathrm{R}}_i$ is the estimated price distribution $\tilde{\mathrm{P}}(\hat{R}_i|\Omega = \omega_i)$ for a single opportunity ω_i. Such distribution $\tilde{\mathrm{R}}_i$ can be estimated, e.g., from histograms of past prices or event outcomes, marginalized over the same or "similar" conditions ($\mathcal{D}_i$) [3]. When a point price estimate is needed, such as when we need to actually trade a selection at a particular price, it is common to take an expected value from $\tilde{\mathrm{R}}_i$, calculated by multiplying each point estimate $\hat{R}_i = \hat{r}_j$ with its associated probability estimation e_j (or probability density $\mathsf{e}(\hat{r})$):

$$\mathbb{E}_{\tilde{\mathrm{R}}_i}[\tilde{\mathrm{e}}(\mathcal{D}_i)] = \sum_j \mathrm{e}_j \cdot \hat{r}_j \quad \text{or alternatively} \tag{2.3}$$

$$\mathbb{E}_{\tilde{\mathrm{R}}_i}[\tilde{\mathrm{e}}(\mathcal{D}_i)] = \int \hat{r} \cdot \mathrm{e}(\hat{r})\ \mathrm{d}\hat{r} \tag{2.4}$$

We note that most of the machine learning models provide directly the point estimates, realizing some functional mapping $\Omega \to [0,1]$.

POINT ESTIMATES The aforementioned market price $m_i \in [0,1]$ of an opportunity ω_i can then be thought of as the market maker's point estimate $m_i = \mathrm{m}(\mathcal{D}_i^m)$ of the fair price r_i. Similarly, the market taker will also try to predict the fair price, which we represent with his/her own point estimate $t_i = \mathbb{E}_{\tilde{\mathrm{T}}_i}[\mathrm{t}(\mathcal{D}_i^t)]$, or simply $t_i = \mathrm{t}(\mathcal{D}_i^t)$. By the definition of his/her role, the market maker m continuously evaluates each opportunity $\forall i : \omega_i \to m_i$. We generally assume that the trader t also has the ability to estimate (predict) the fair price $\forall i : \omega_i \to t_i$ of each opportunity in the market.[9] The trader t then must posses his/her own estimate, i.e. be generally different from m[10], and the fair price r must be unknown at the time of trading, otherwise there would be no incentive to trade.

Definition 2.5.1. We can now generalize the reasoning about individual opportunities and estimates to reasoning over the whole joint *market distribution* $\mathrm{P}_\Omega(R, M, T)$ capturing the relationships between the estimates across a whole set of opportunities $\omega_1, \ldots, \omega_n \in \Omega$ generated by the market environment. For each such opportunity ω_i, we will thus operate with 3 distinct values we will refer to as (i) the fair price $R = r_i$, (ii) the market maker's estimate $M = m_i$ and (iii) the market taker's estimate $T = t_i$.

Given the uncertainty, it follows that both the m_i and t_i estimates are always going to be to some extent erroneous, where the former guaranties existence of mispriced assets. Note that this is a necessary but insufficient condition for market inefficiencies to exist (Section 2.2).

OPTIMIZATION Typically, one optimizes the estimator by tuning its parameters $\theta \in \Theta$ to fit some historical data $\mathcal{D}^t$ relevant for the prediction of the underlying fair prices R. Estimation of unknown parameters θ from empirical data $\mathcal{D}$ is then one of the key problems in statistics [44]. There are several views on the problem. One class of approaches is to maximize probability of the observed data with methods such as maximum likelihood $\mathrm{p}(\mathcal{D}|\theta)$ or maximum a-posteriori $\mathrm{p}(\theta|\mathcal{D})$ estimation. One can also directly search for an estimator with some desired target properties, such as minimum-variance unbiased estimator [130] or best linear unbiased estimator [54]. A

[9] Alternatively, a systematically selected subset can be considered instead.

[10] since e.g. copying the estimate from the market maker would not lead to any trading incentive.

very common methodology is Bayesian estimation, both with or without an informative prior, where one tries to minimize (posterior) expectation of some error function [11].

ESTIMATION ERROR The quality of an estimator e can then be expressed through its empirical error $err(\mathsf{e})$ over some set of opportunities Ω. Since we assume the role of the trader t, we further present error measurements between the fair prices R and the trader's estimates T. Some of the most popular error measures then include the mean square error (MSE):

$$MSE_{\Omega}(R,T) = \mathbb{E}[(T-R)^2] = \frac{1}{|\Omega|}\sum_{\omega_i \in \Omega}(t_i - r_i)^2 \tag{2.5}$$

and the mean crossentropy ($XENT$):

$$XENT_{\Omega}(R,T) = \mathbb{E}_R[-log(T)] = \frac{1}{|\Omega|}\sum_{\omega_i \in \Omega}\sum_{j \in outcomes} r_j \cdot log(t_j) \tag{2.6}$$

Note that these correspond to the market sides ζ. In the case of the two-selection markets corresponding to binary event outcomes, we can rewrite $XENT$ as

$$XENT_{\Omega}(R,T) = \frac{1}{|\Omega|}\sum_{\omega_i \in \Omega} r_i \cdot log(t_i) + (1-r_i) \cdot log(1-t_i) \tag{2.7}$$

which can then also be also understood as a regression of the underlying value r_i of each opportunity ω_i.

These particular error functions are of special interest as minimizing $XENT$ generally corresponds to maximizing the data (log-)likelihood, while MSE corresponds to maximizing the data likelihood with a linear gaussian model [74]. The crossentropy error is then also closely linked to a common measure of "distance" between probability distributions known as Kullback-Leibler divergence [69], which is defined as

$$D_{KL}(R||T) = -\sum_i r_i \cdot log\frac{t_i}{r_i} \tag{2.8}$$

since

$$XENT(R,T) = H(R) + D_{KL}(T||R) \tag{2.9}$$

where $H(R)$ is the entropy of the fair price distribution R. Considering that true distribution being estimated does not change, its entropy $H(R)$ can be considered a constant, rendering the cross-entropy error $XENT(R,T)$ minimization equivalent to minimizing the the KL-divergence $D_{KL}(R||T)$, sometimes also referred to as the relative entropy [11] (see Section 16.1 for further connections).

2.6 EXPECTED RETURNS

Being able to predict the future price in a prediction market can be directly turned into positive returns in trading. In the prediction markets, the relative frequency of the observed outcomes will approach the fair price in the long run. Being able to correctly estimate the price guaranties systematic profits, even though actual profits might deviate from the expectation in short term.

Since the total profit is dependent on the actual amount invested, which is yet to be determined by the investment strategy (Section 2.7), the traders commonly assess *relative* profitability of individual opportunities through measures such return on investment (RoI), which we further denote as ρ_i. In general, ρ_i is simply the relative return made from an investment.

The uncertain, stochastic nature of the prediction problem renders the value estimates $\hat{r}_i$ for a particular ω_i as random variables $\hat{r}_i \sim \tilde{\mathrm{R}}_i$ (Section 2.5). Consequently, a return ρ_i derived from such an estimate $\hat{r}_i$ will be a random variable, too. Instead of the actual ρ_i we can thus again calculate with its expectation $\mathbb{E}[\rho_i]$ w.r.t. the used estimator e, further denoted as $\mathbb{E}_{\mathsf{e}}[\rho_i]$.

The market price (odds) directly reflect the relative returns, and so the expected ρ_i of an opportunity ω_i based on a probability estimated by t can be calculated as

$$\mathbb{E}_{\mathsf{t}}[\rho_i^\alpha] = \mathbb{E}_{\tilde{T}_i}\left[\frac{\tilde{T}_i}{M(\omega_i^\beta)} - 1\right] = \frac{t_i}{m_i{}^\beta} - 1 \tag{2.10}$$

$$\mathbb{E}[\rho_i^\beta] = \mathbb{E}_{\tilde{T}_i}\left[\frac{\tilde{T}_i}{M(\omega_i^\alpha)} - 1\right] = \frac{1-t_i}{1-m_i{}^\alpha} - 1 \tag{2.11}$$

For example, a bet of \$100 on an outcome with estimated probability $t_i = \frac{3}{4}$, and associated bookmaker's (decimal) odds of $\frac{1}{M(\omega_i)} = 2.0$, i.e. yielding a net return of \$100 if realized, and a loss of $\$-100$ if not, will also result into a ROI of $\rho_i = 0.5$ (50%).

Although the average returns should converge to the true expected returns in the long run, these can still be very different from the predicted expected returns since generally $\mathbb{E}_{\mathsf{t}}(\rho_i) \neq \mathbb{E}_{\mathsf{r}}(\rho_i)$. The discrepancy is of course conditioned by the quality of the predictor t w.r.t. the true r. The approach to minimize the prediction error (Section 2.5) then seems very natural, and there are also some theoretical guaranties like, for instance, in the case where $XENT(R,T) < XENT(R,M)$, i.e. the investor possesses a price prediction model superior to the market maker in terms of cross-entropy, we are guaranteed to make long-term profits with optimal investment routines such as the Kelly strategy (Section 2.7).

RISK AND UTILITY To explicitly account for the discussed uncertainty involved in trading, the concept of *risk* assessment has been proposed. This means that instead of direct maximization of the expected returns, one should strive for a balance between the expected profit and risk, stemming from the uncertainty. The risk can then be quantified by statistical means such as the variance of the expected profit [85] or probability of a drawdown [21]. Note that apart from the quantifiable risk, there is also a structural risk stemming from the fact that, similarly to the expected returns, the assessment of risk is based on merely estimated parameters.

Not all investors then share the same preferences to balance the expected profit and risk in the same manner. To incorporate individual preferences into the decision making, the concept of a *utility function* u has been proposed to steer the investment optimization process. A utility function generally maps each alternative onto a real number, defining a total ordering over some set of alternative investments. In our case it is some monotonically growing function u transforming the net returns ρ into a new real quantity $\mathsf{u}(\rho)$ to be optimized.[11] The concept of risk is then closely connected to utility, as maximizing any concave utility function directly reflects a preference for risk aversion [6].

2.7 INVESTMENT STRATEGIES

An investment strategy can be seen as the final step of the traders's workflow. Given the expected returns from individual trading opportunities present at a time, the trader needs to decide on how to *allocate* his/her wealth across the available opportunities Ω in order to optimize his/her utility u, i.e. some requested trade-off between expected returns and risk. Formally, the investment strategy is a function s mapping a vector of opportunities $\boldsymbol{\omega}$ onto a wealth allocation vector $\boldsymbol{f}$:

$$\mathsf{s} : \boldsymbol{\omega} \mapsto \boldsymbol{f}. \tag{2.12}$$

The vector of the wealth allocations $\boldsymbol{f}$, corresponding to portions of some current wealth W, is then often referred to as a *portfolio* over the opportunities $\boldsymbol{\omega}$. We note that the trader might also want to leave certain fraction of the wealth W aside, which is commonly incorporated by introducing an extra opportunity ("cash option") with a constant zero[12] net return. Generally, there are several approaches to the wealth allocation problem and we will briefly review some of the most popular ones [78].

2.7.1 *Unit Stake*

The most trivial investment strategy a unit-staking strategy where one independently allocates the same absolute amount d on every opportunity with an assumed positive expected return:

[11] Given the stochastic setting, we will again consider expected utility $\mathbb{E}[\mathsf{u}(\rho)]$ instead.

[12] This might also be further extended by introducing a slightly positive net return instead to emulate the option of, e.g., storing the money in a bank account with an interest.

$$\mathsf{s} : \rho_i \mapsto \begin{cases} f_i = d, & \text{if } \mathbb{E}_\mathsf{t}[\rho_i] > 0 \\ f_i = 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

Despite being very naive, this strategy is also robust against estimation errors [104], since the allocation simply remains constant no matter the circumstances. Note that the individual investments d are not considered as fractions relative to the current wealth W here, and so a small enough unit $d \ll \mathsf{W}$ has to be chosen so that it is possible to invest into all profitable opportunities. Given that the allocated unit d is small enough, this strategy will typically also have a very conservative risk profile, nevertheless the expected portfolio profits can be way below optimal. The size of d then remains a hyperparameter the choice of which is left to the user.

2.7.2 *Modern Portfolio Theory*

A more principled approach is that of the Modern Portfolio Theory (MPT) [85] which strives to balance optimally between the expected return and risk. The general idea behind MPT is that a portfolio $\boldsymbol{f}^1$, i.e. a vector of asset capital allocations $\boldsymbol{f} = f_1, \ldots, f_n$ over some opportunities $\omega_1, \ldots, \omega_n$, is superior to $\boldsymbol{f}^2$, if its corresponding expected return ρ (Section 2.6) is at least as great

$$\mathbb{E}_\mathsf{t}[\boldsymbol{\rho} \cdot \boldsymbol{f}^1] \geq \mathbb{E}_\mathsf{t}[\boldsymbol{\rho} \cdot \boldsymbol{f}^2] \tag{2.14}$$

and a given risk measure $risk : \mathbb{R}^n \to \mathbb{R}$ of the portfolio w.r.t. the returns is no greater

$$risk_{\mathbb{E}_\mathsf{t}[\rho]}\left(\boldsymbol{f}^1\right) \leq risk_{\mathbb{E}_\mathsf{t}[\rho]}\left(\boldsymbol{f}^2\right). \tag{2.15}$$

This creates a partial ordering on the set of all possible portfolios. When combined into a joint utility, we can trade-off the expected profit vs. risk by maximizing the following

$$\underset{\boldsymbol{f} \in \mathbb{R}^n}{\text{maximize}} \left(\mathbb{E}_\mathsf{t}[\boldsymbol{\rho} \cdot \boldsymbol{f}] - \gamma \cdot risk_{\mathbb{E}_\mathsf{t}[\rho]}(\boldsymbol{f})\right), \tag{2.16}$$

where γ is a hyperparameter reflecting the user's preference for risk.

In the most common setup, the *risk* of a portfolio $\boldsymbol{f}$ is measured through the expected total variance of its profit $\text{Var}[\boldsymbol{\rho} \cdot \boldsymbol{f}] = \boldsymbol{f}^T \Sigma \boldsymbol{f}$, based on a given covariance matrix $\boldsymbol{\Sigma}_n^n$ of returns of the individual opportunities, which can be again estimated from historical data (Section 2.4). MPT can then be expressed as the following constrained maximization problem:

$$\begin{aligned} \underset{\boldsymbol{f} \in \mathbb{R}^n}{\text{maximize}} \quad & \mathbb{E}_\mathsf{t}[\boldsymbol{\rho} \cdot \boldsymbol{f}] - \gamma \cdot \boldsymbol{f}^T \Sigma \boldsymbol{f} \\ \text{subject to} \quad & \sum_{i=1}^{n} f_i = 1 \end{aligned} \tag{2.17}$$

Note that the capital allocations sum up to one as they simply reflect fractions of the current bankroll W (including the fraction left in "cash").

The main weakness of MPT is that the variance of profit is hardly a good measure of risk for profit distributions other than Gaussian [109]. Apart from the variance $\mathrm{Var}[\boldsymbol{w}]$ of the potential net returns $\boldsymbol{w} = \boldsymbol{\rho} \cdot \boldsymbol{f}$, different risk measures have been proposed [85], such as standard deviation $\sigma(\boldsymbol{w}) = \sqrt{\mathrm{Var}[\boldsymbol{w}]}$ and coefficient of variation $CV(\boldsymbol{w}) = \frac{\sigma(\boldsymbol{w})}{\mathbb{E}[\boldsymbol{w}]}$. Nevertheless these all share the same weakness. Generally, there is no agreed-upon measure of risk, rendering the whole concept a bit dubious. Moreover, the strategy only works with the opportunities $\boldsymbol{\omega}$ currently at hand, and thus ignores any knowledge about the actual market distribution P_Ω.

SHARPE RATIO Apart from the choice of the risk measure, the inherent degree of freedom in MPT is how to select a particular portfolio from the efficient frontier (based on the choice of γ). Perhaps the most popular way to avoid the dilemma is to select a spot in the pareto-front with the highest expected profits w.r.t. the risk. For the risk measure of $\sigma(\boldsymbol{w})$, this is known as the "Sharpe ratio" [114], generally defined as

$$\frac{\mathbb{E}_{\mathrm{t}}[\boldsymbol{w}] - r_f}{\sigma(\boldsymbol{w})}, \tag{2.18}$$

where $\mathbb{E}[\boldsymbol{w}]$ is the expected return of the portfolio, $\sigma(\boldsymbol{w})$ is the standard deviation of the return, and r_f is a "risk-free rate". We do not consider any risk free investment in our setting, and so we can reformulate the optimization problem as

$$\begin{aligned} \underset{\boldsymbol{f} \in \mathbb{R}^n}{\text{maximize}} \quad & \frac{\mathbb{E}_{\mathrm{t}}[\boldsymbol{\rho} \cdot \boldsymbol{f}]}{\sqrt{\boldsymbol{f}^T \boldsymbol{\Sigma} \boldsymbol{f}}} \\ \text{subject to} \quad & \sum_{i=1}^{n} f_i = 1 \end{aligned} \tag{2.19}$$

2.7.3 *Kelly Criterion*

The Kelly criterion [63, 124] assumes the investment problem in time[13], i.e. it optimizes multi-period investments in contrast to MPT which is concerned only with single-period portfolio returns. It is based on the idea of expected multiplicative growth W_{G} of a continuously reinvested bankroll W_τ. The goal is to a find a portfolio $\boldsymbol{f}$ such that the long-term expected value of the resulting profit $\mathrm{W}_{\tau \to \infty}$ is maximal, which is equivalent to maximizing the geometric growth rate of wealth defined as

$$W_{\mathrm{G}} = \lim_{t \to \infty} log \left(\frac{W_t}{W_o} \right)^{\frac{1}{t}}. \tag{2.20}$$

[13] Note this is in contrast to MPT which assumes the problem in an ensemble of traders at the same time, i.e. through expectation.

For its multiplicative nature, it is also known as the geometric mean policy, emphasizing the contrast to the arithmetic mean approaches (e.g. MPT) based directly on the expected value of wealth. The two can, however, be looked at similarly with the use of a logarithmic utility function, transforming the geometric into the arithmetic mean, and the expected geometric growth rate into the expected value of wealth, respectively. The problem can then be again expressed by the standard means of maximizing the (estimated) expected utility value as

$$\begin{aligned} \underset{\boldsymbol{f} \in \mathbb{R}^n}{\text{maximize}} \quad & \mathbb{E}_t \left[\log \left(1 + \boldsymbol{f}^T \cdot \boldsymbol{\rho} \right) \right] \\ \text{subject to} \quad & \sum_{i=1}^{n} f_i = 1 \end{aligned} \tag{2.21}$$

Note that, in contrast to MPT, there is no explicit term for risk here, as the notion of risk is inherently encompassed in the growth-based view of the wealth progression, i.e. the long-term value of a portfolio that is too risky will be smaller than that of a portfolio with the right risk balance (and similarly for portfolios that are too conservative). The risk is thus captured by the logarithmic (concave) utility transformation itself.

The calculated portfolio is then provably optimal, i.e. it accumulates more wealth than any other portfolio chosen by any other strategy in the long-run. However, this strong result only holds given, considerably unrealistic, assumptions [63, 103, 124]. Similarly to MPT, we assume to know the true returns while calculating merely with estimates and additionally, given the underlying growth perspective, that we are repeatedly presented with the same opportunities from P_Ω ad infinitum, making the optimality of the growth-based risk treatment in Kelly likewise a bit dubious. Despite the fact that the given conditions are impossible to meet in practice, the Kelly strategy is very popular, particularly its various modifications to mitigate the aforementioned issues.

FRACTIONAL KELLY The result of the Kelly optimization problem is, for each opportunity, the ideal fraction $\omega \mapsto f^*$ one is ought to invest to achieve the maximal long-term profits. The fraction f^* thus dictates an upper-bound on the possible profit, meaning that increasing the invested fraction further will actually decrease the long-term profit.[14] This is commonly known as "overbetting". Since the true expected return ρ is unknown, however, such a situation might occur even while betting with a fraction assumed to be optimal. Intentionally decreasing the calculated fraction f^* by some ratio $\frac{1}{d'}$ then decreases the risk of overbetting stemming from a possibly overvalued estimate. Such an approach is commonly referred to as "fractional Kelly" [83]. Ideally, one should estimate the optimal shrinkage d' as another hyperparameter [8, 127] based

[14] this is due to the assumed multiplicative, growth-based view of Kelly, which is in contrast to the additive MPT, where overbetting would merely increase the risk.

on backtesting performance, however, it is very common to simply choose a fixed ratio such as $\frac{1}{2}$ of the estimated optimal Kelly fraction f^*, commonly referred to as "half Kelly" by practitioners. While there are other remedies to mitigate the risk with the Kelly criterion [21, 123], fractional Kelly is a very effective method which is widely adopted in practice due to its simplicity. In addition to mitigating the overbetting risk, it generally decreases volatility, which also tends to be considerably high with the plain Kelly criterion.

CORRESPONDENCE TO MPT While Kelly is clearly based on different principles than MPT, there is an interesting close connection between the two strategies. Following [21], let us make an assumption for a Taylor series approximation that our net profits are not too far from zero $\boldsymbol{\rho}^T \cdot \boldsymbol{f} \approx \mathbf{0}$, allowing us to proceed with the Taylor expansion of the optimized growth as

$$\log\left(1 + \boldsymbol{\rho}^T \cdot \boldsymbol{f}\right) = \boldsymbol{\rho}^T \cdot \boldsymbol{f} - \frac{\left(\boldsymbol{\rho}^T \cdot \boldsymbol{f}\right)^2}{2} + \dots \tag{2.22}$$

Now taking only the first two terms from the series we transform the expectation of logarithm into a new problem objective as follows

$$\underset{\boldsymbol{f} \in \mathbb{R}^n}{\text{maximize}} \quad \mathbb{E}\left[\boldsymbol{\rho}^T \cdot \boldsymbol{f} - \frac{\left(\boldsymbol{\rho}^T \cdot \boldsymbol{f}\right)^2}{2}\right] \tag{2.23}$$

Note that, interestingly, the problem can now be rewritten to

$$\begin{aligned} &\underset{\boldsymbol{f} \in \mathbb{R}^n}{\text{maximize}} && \mathbb{E}\left[\boldsymbol{\rho}^T \cdot \boldsymbol{f}\right] - \frac{1}{2}\mathbb{E}\left[\boldsymbol{f}^T\left(\boldsymbol{\rho} \cdot \boldsymbol{\rho}^T\right)\boldsymbol{f}\right] \\ &\text{subject to} && \sum_{i=1}^{n} f_i = 1 \end{aligned} \tag{2.24}$$

corresponding to the original MPT formulation from Equation 2.17 for the particular user choice of $\gamma = \frac{1}{2}$. It follows from the fact that the geometric mean is approximately the arithmetic mean minus $\frac{1}{2}$ of variance [85], providing further insight into the connection of the two popular strategies of Kelly and Markowitz, respectively. While the solution is merely an approximation, it also tends to be more robust to estimation errors than the original Kelly, similarly to the fractional approach.

Part II

SCORE-BASED MODELING: EXPERIMENTAL REVIEW

3

INTRODUCTION

In Section 2.5, we introduced the notion of price estimators. Naturally, many different models have already been developed to estimate the probabilities (prices) for different markets. Unsurprisingly, most of the models focus on the Moneyline market (i. e.winner of a match). As we aspire to develop our own models, we must first explore the state-of-the-art. The body of the literature is very diverse, spanning across multiple domains. However, one of the domains is more prevalent than the others.

Soccer, being arguably the most popular sport in the world, continues to attract researchers and practitioners competing for the design of the most accurate game outcome forecasting models. However, due to a lack of a standardized dataset, it has been difficult to draw conclusive statements about the relative performances of the diverse approaches. The creation of such a dataset has been, however, further complicated by the fact that the proposed models often utilize varying details of match and background information in order to gain more advantage over the competition.

While for some of the top leagues, complete information about the game, including player-tracking data, can be obtained, such an approach does not generalize onto the vast amount of the lower leagues, where merely the results with basic metadata are all that is being stored for each match. Moreover, the fine-grained data are often proprietary and rather expensive, rendering them unsuitable for use in academic benchmarks.

For the purpose of a sound experimental comparison, we propose to target the broadest possible scope of the domain by considering solely the *score-based* models, i. e., the models that use the final scores, teams' names and dates as the only input covariates. Such an elementary modeling paradigm allows us to calculate predictions for virtually all existing matches and, consequently, unify the training and evaluation protocol across the diverse approaches.

Conveniently, a large dataset containing 218,916 match results from 52 leagues since the season 2000/01 was released by Dubitzky et al. [34]. The records in the dataset consist merely of the league names, dates, team names, and the resulting scores. The availability of such a large dataset provides an ideal opportunity to finally shed some light on the relative performance of the respective score-based methods. For that purpose, we have reimplemented the most promising models from the literature to analyze their performance under a unified protocol.

The rest of this part is organized as follows. In Section 4, we summarize the relevant research. Section 5 provides a brief description of the implemented models. Section 6 explains the protocol for fitting and evaluation of the models. Experimental results are compiled in Section 7, and we conclude the part in Section 8.

4

RELATED WORK

The body of related work on score-based predictive models can be generally divided into 3 categories: (i) *statistical models*, where the goals scored are assumed to follow a particular parametric probability distribution, (ii) *rating systems* that assign a real-valued rating(s) to each team to capture its strength, and (iii) *machine learning models* where various complex features are usually derived from the data and passed to an off-the-shelf learning algorithm.

4.1 STATISTICAL MODELS

The research in the domain of score-based soccer modeling has traditionally been dominated by statistical approaches. In his pioneering work, Maher [84] came up with a double Poison model and bivariate Poisson model, where the latter provided a better fit for the data. Maher also introduced the notion of teams' attacking and defensive strengths and how to use them for forecasting of the match results. This notion is still used in the current research nowadays.

Dixon and Coles [33] extended Maher's ideas, as they introduced a dependency between the home and away teams' goals scored for the double Poisson model, effectively increasing the probabilities of low-scoring draws. Also, while Maher considered the strength of the team to be time-invariant, here the idea of weighting the likelihood during fitting of the model was introduced. Particularly, the authors used exponential time weighting to discount the effects of past results. The simplicity of exponential time weighting allows for its use with other models too [77]. A different approach to the time evolution of teams was used in Rue and Salvesen [110], where the authors used a Brownian motion to tie together the teams' strength parameters in consecutive rounds. Crowder et al. [30] used an autoregressive model for the evolution of teams' strengths, improving on results by Dixon and Coles [33] and on the computational complexity of Rue and Salvesen [110]. A static hierarchical model based on the double Poisson distribution was introduced by Baio and Blangiardo [7], claiming performance non-inferior to the bivariate Poisson model [61]. Owen [97] used a random walk to model the teams strength evolution in the double Poisson model, however a comparison against the established likelihood weighting approach was not done. Koopman and Lit [66] introduced time dynamics into the bivariate Poisson model using a state-space model representation. The authors also pointed out that the dependency between scores had little effect on the out-of-sample forecasting performance of the model. This observation was latter supported by Ley, Wiele, and Eetvelde [77]. Angelini and De Angelis [4] investigated another technique for implementing the time-dynamics with a PARX model [1]. The PARX model outperformed Dixon and Coles [33] in forecasting the number of scored goals. Koopman and Lit [67] compared bivariate Poisson, Skellam, and ordered probit models where the teams' strengths were updated according to a time series model. The bivariate Poisson model achieved the best results.

Karlis and Ntzoufras [61] noticed that the bivariate Poisson models tend to underestimate the probabilities of draws and introduced a diagonal-inflated bivariate Poisson model. Karlis and Ntzoufras [62] then eliminated the need to explicitly model the scores dependency by using the Skellam distribution [116], where the evolution of the teams' strengths was implemented using Bayesian updates. McHale and Scarf [88] experimented with negative binomial and bivariate Poisson models where the dependence structure was implemented using copulas. The most recent novelty in statistical approaches is the use of bivariate Weibull count model [17]. Unlike in the Poisson distribution, where the mean is equal to the variance, the Weibull count distribution is determined by two parameters, allowing for better handling of both under and over-dispersed data. The bivariate model is constructed using a copula function. The model provides a better fit for the data than the Poisson model at the expense of higher computational time, as the calculation of the probability density function of the Weibull count model is very demanding. A great review of the statistical approaches can be found in Ley, Wiele, and Eetvelde [77].

4.2 RATING SYSTEMS

Another technique to estimate the strength of an individual or a team are the so-called rating systems. Ratings try to capture the team's strength into one or two scalar values, providing relative ordering of the teams, but not necessarily a way to obtain the probabilistic forecasts. The world's best-known rating system is the Elo rating [35], originally used for assessing the strength of chess players. The player's performance is assumed to be drawn from a Gaussian distribution with a fixed variance. The mean of such distribution is then the player's rating (skill). An application of the Elo rating in the domain of soccer was shown in Hvattum and Arntzen [57]. Recent work by Robberechts and Davis [108] demonstrated that the method yields competitive results.

TrueSkill [49] is another system that enhances the Elo rating by operating not only with the variance of the player's skill (rating) but also with the variance of his/her performance. This variance reflects the uncertainty about the player's skill in situations when we have not yet observed enough data (performances). The authors demonstrated faster convergence and better predictive performance in comparison with the Elo rating.

One of the caveats of the TrueSkill is that it does not propagate the newly obtained information backward to correct the past ratings. In other words, the method does filtering without smoothing. The work by Dangauthier et al. [32] aimed to fix this issue. Also, the plain version of TrueSkill does not account for the score difference, as it only considers the ternary win-draw-loss outcome of a match. Guo et al. [52] proposed an extension to take into account the score differences and claimed superior performance to the vanilla TrueSkill, also on a soccer dataset. The current evolution of the TrueSkill rating system is TrueSkill2 [93], however most of the improvements are domain-specific to matchmaking in online games, which is the primary focus of the system.

A soccer domain-specific rating system called pi-ratings was introduced in Constantinou and Fenton [27]. The team's strength is represented by its home and away ratings, which are updated after each match according to manually set learning rates. Another score-based rating system was developed by Berrar, Lopes, and Dubitzky [13], where the rating system parameters were tuned using particle swarm optimization and fed to a standard off-the-shelf learner. A method for ranking teams after an incomplete season was proposed by [31].

4.3 MACHINE LEARNING APPROACHES

Machine learning models are not very common in score-based modeling as they usually leverage extra features besides the scores or ratings. Constantinou [25] extended his pi-ratings model with a Bayesian network to obtain the probability distribution over possible match outcomes from the rating difference. Tsokos et al. [125] tested several variations of the Bradley-Terry model [9, 18] and a hierarchical Poisson model. In the end, the hierarchical Poisson model outperformed all the Bradley-Terry models. The inferiority of the Bradley-Terry model to other methods was further confirmed by Ley, Wiele, and Eetvelde [77].

A different, unorthodox approach to the problem is to view the match data as a relational structure (graph) between the teams. This was first pointed out by Van Haaren and Broeck [129] where the authors achieved some promising results. The graph representation of the data was also utilized by Govan, Meyer, and Albright [48], who used the PageRank algorithm [98] to estimate the teams' strengths. The same author later proposed a so-called offense-defense model [47], which can be seen as an analogy to the HITS algorithm [65].

5

REVIEWED MODELS

Our ambition is to provide an experimental comparison of a variety of different approaches towards the problem of soccer match outcome prediction. For that goal, we have reimplemented and tested the most prevalent models in score-based soccer forecasting, as well as some models that, despite their promising results, have not received as much attention as the former. The selected models, together with the underlying reasoning, are then as follows.

The double Poisson model with exponential time weighting [33, 84] is probably the most established model to date. Recent work by Ley, Wiele, and Eetvelde [77] showed that the model (and its bivariate variant) is still relevant nowadays. The most notable improvement upon these models was claimed by Boshnakov, Kharrat, and McHale [17] using their bivariate Weibull model.

From the perspective of the rating systems, the Elo [35, 57] proved to be competitive by Robberechts and Davis [108]. Constantinou and Fenton [27] claimed to outperform the Elo considerably. Another rating system that claimed improvement over the Elo, was TrueSkill [49]. Its extension for score-based forecasting by Guo et al. [52] demonstrated better results on a soccer dataset, therefore we chose the extension over the original model. Steph ratings [121] not only did well in a Kaggle competition, but they are extension of another successful rating system – the Glicko [45]. Recently, the ratings by Berrar, Lopes, and Dubitzky [13] showed the most promising results.

5.1 STATISTICAL MODELS

In this section we take a closer look at the statistical models included in this review. As all the statistical models provide a likelihood function for the scores, the probability of a team winning/drawing/losing can be easily computed by marginalizing the probability distribution over the results.

5.1.1 *Double Poisson Model*

Double Poisson represents one of the earliest [84] and simplest models. However, as was shown in Ley, Wiele, and Eetvelde [77], it still remains very competitive. The model assumes the goals scored by the competing teams in a match to be independent. Therefore, the probability of a home team scoring x goals with the away team scoring y goals is given by

$$P(G_H = x, G_A = y|\lambda_H, \lambda_A) = \frac{\lambda_H^x e^{-\lambda_H}}{x!} \cdot \frac{\lambda_A^y e^{-\lambda_A}}{y!}. \tag{5.1}$$

where λ_H and λ_A are the scoring rates of the teams (the means of the underlying Poisson distributions). The scoring rates of the teams for a particular match can be expressed in terms of the Maher's specification as

$$\begin{aligned} \log(\lambda_H) &= Att_H - Def_A + H \\ \log(\lambda_A) &= Att_A - Def_H \end{aligned} \tag{5.2}$$

where H represents a home advantage, and Att and Def are respectively the offensive and defensive strengths of the teams (the actual model parameters).

Later, Ley, Wiele, and Eetvelde [77] demonstrated that the number of the model's parameters can be effectively halved by considering only a single strength parameter for each team without any loss of predictive performance, i.e. reducing to

$$\begin{aligned} \log(\lambda_H) &= Str_H - Str_A + H \\ \log(\lambda_A) &= Str_A - Str_H \end{aligned} \tag{5.3}$$

5.1.2 *Bivariate Poisson Model*

Karlis and Ntzoufras [61] extended the double Poisson model by introducing dependence between the scored and conceded goals. The dependence is given by another Poisson distribution. The scored goals are modelled as $G_H = X_1 + X_3$, $G_A = X_2 + X_3$ where $X_1 \sim Pois(\lambda_H)$, $X_2 \sim Pois(\lambda_A)$ and $X_3 \sim Pois(\lambda_C)$. The probability function for the bivariate distribution is then given by

$$\begin{aligned} &P(G_H = x, G_A = y|\lambda_H, \lambda_A, \lambda_C) = \\ &= e^{-(\lambda_H+\lambda_A+\lambda_C)} \frac{\lambda_H^x \lambda_A^y}{x!y!} \sum_{k=0}^{min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_C}{\lambda_H \lambda_A}\right)^k \end{aligned} \tag{5.4}$$

where the scoring rates λ_H and λ_A are computed in the same fashion as for the double Poisson model (Eq. 5.3), and $log\lambda_C$ is fitted together with the Str and H parameters.

5.1.3 *Double Weibull Count Model*

One of the pitfalls of the Poisson-based models is that the Poisson distribution does not consider under or over-dispersion in the data since the variance of the distribution is strictly equal to the mean. Weibull-based models [17] aim to tackle this issue. The Weibull count model was derived from the continuous Weibull distribution by McShane et al. [89]. The probability density function of the univariate Weibull count model is given by

$$P(G = x|\lambda, c) = \sum_{j=x}^{\infty} \frac{(-1)^{x+j}\lambda\alpha_j^x}{\Gamma(cj+1)}, \tag{5.5}$$

where c is the shape parameter of the distribution, and α_j^x is defined recursively for $x = 0, 1, \ldots$ and $j = x+1, x+2, \ldots$ as

$$\alpha_j^0 = \frac{\Gamma(cj+1)}{\Gamma(c+1)} \tag{5.6}$$

$$\alpha_j^{x+1} = \sum_{m=x}^{j-1} \alpha_m^x \frac{\Gamma(cj-cm+1)}{\Gamma(c-j+1)} \tag{5.7}$$

The probability of observing a score in a soccer match is then obtained by multiplying the two probability distributions for each of the opposing teams, analogically to Eq. 5.4. For this reason, the double Poisson and Weibull models are also referred to as "independent". Calculation of the scoring rates λ then also follows Eq. 5.3.

5.1.4 *Bivariate Weibull Count Model*

The bivariate version of the Weibull count model was introduced by Boshnakov, Kharrat, and McHale [17]. The Weibull marginals were tied together with Frank copula to form the bivariate model. The joint probability function is given by

$$\begin{aligned}
P(G_H = x, G_A = y|\lambda_H, \lambda_A, c_H, c_A) = \\
&= C\left(F\left(x|\lambda_H, c_H\right), F\left(y|\lambda_A, c_A\right)\right) \\
&- C\left(F\left(x-1|\lambda_H, c_H\right), F\left(y|\lambda_A, c_A\right)\right) \\
&- C\left(F\left(x|\lambda_H, c_H\right), F\left(y-1|\lambda_A, c_A\right)\right) \\
&+ C\left(F\left(x-1|\lambda_H, c_H\right), F\left(y-1|\lambda_A, c_A\right)\right)
\end{aligned} \tag{5.8}$$

where F is a cumulative distribution function that can be computed using the probability density function from Eq. (5.5) and c_H, c_A are the shape parameters of the distribution. The C is Frank copula function given by:

$$C(u, v) = -\frac{1}{\kappa}\ln\left(1 + \frac{(e^{-\kappa u - 1})(e^{-\kappa v - 1})}{e^{-\kappa} - 1}\right), \tag{5.9}$$

where κ is the dependence parameter. Calculation of the scoring rates λ again follows Eq. (5.3).

5.2 RATING SYSTEMS

One of the main differences between statistical models and rating systems is that rating systems were designed mainly to rank the competing teams in a league and not necessarily to produce a probability distribution over the possible outcomes. However, this can be effectively solved by employing a subsequent regression model that transforms the ratings into the desired probability distribution, as was demonstrated in Hvattum and Arntzen [57]. The details on how the ratings and the subsequent regression are trained can be found in Section 6.1.

5.2.1 *Elo Ratings*

Elo [35] is a general rating system, the modification of which is still used for evaluation of the strength of chess players. Hvattum and Arntzen [57] proposed its modification for soccer and consequently, Robberechts and Davis [108] demonstrated the effectiveness of the method. The modification involves the use of an ordered logit model [87] to obtain the probability distribution over the possible match outcomes. At the core, each team's performance is assumed to be normally distributed around its true strength. The expected outcomes for both teams are then calculated as follows:

$$E_H = \frac{1}{1 + c^{(R_A - R_H)/d}} \tag{5.10}$$

$$E_A = 1 - E_H \tag{5.11}$$

where R_H and R_A are the ratings of the home and away teams, and c and d are metaparameters of the method. The actual ternary outcome of the match is then encoded numerically as

$$S_H = \begin{cases} 1 & \text{if the home team won} \\ 0.5 & \text{if the match was drawn} \\ 0 & \text{if the home team lost} \end{cases} \tag{5.12}$$

Finally, after the match, the ratings of the teams are updated using

$$R_H^{t+1} = R_H^t + k(1+\delta)^\gamma \cdot (S_H - E_H) \tag{5.13}$$

$$R_A^{t+1} = R_A^t - k(1+\delta)^\gamma \cdot (S_H - E_H) \tag{5.14}$$

where δ is an absolute value of goal difference, k represent a learning rate and γ is a metaparameter scaling the influence of the goal difference on the rating change.

5.2.2 *Steph Ratings*

Steph ratings [121] are an evolution of another rating system known as Glicko[45] developed for a chess rating competition at Kaggle. However, the ratings can be easily adapted to other sports. The strength of a team is represented with a tuple (r, v) to capture the team's rating and its variance. Unlike in Elo, the variance of the rating is not constant. Before each match, the variance is increased based on the time passed (Δt) since the last match of the team and a scaling factor c

$$v^t \mathrel{+}= c\Delta t. \tag{5.15}$$

The expected outcome (e) is then computed, accounting for the rating difference (Δr) between the competing teams (i and j), and a home advantage (γ).

$$w = \begin{cases} -1 & \text{for home team} \\ 1 & \text{for away team} \end{cases} \tag{5.16}$$

$$q = \frac{\log 10}{400} \tag{5.17}$$

$$\Delta r = w \cdot (r_i^t - r_j^t + \gamma) \tag{5.18}$$

$$k = \frac{1}{1 + 3q^2 v_j^t / \pi^2} \tag{5.19}$$

$$e = \frac{1}{10^{-k\Delta r/400}} \tag{5.20}$$

Finally, the rating and its variance is updated followingly:

$$s = \begin{cases} 1 & \text{if team won} \\ 0.5 & \text{in case of draw} \\ 0 & \text{if team lost} \end{cases} \tag{5.21}$$

$$d = q^2 k^2 e(1 - e) \tag{5.22}$$

$$v_i^{t+1} = \left(\frac{1}{v_i^t + h} + d \right)^{-1} \tag{5.23}$$

$$r_i^{t+1} = r_i^t + qv^{t+1}k(s - e + b) + \lambda(r_j^t - r_i^t) \tag{5.24}$$

where λ is scaling factor for rating difference and h controls the increase in rating's variance over time. The b serves as a bonus to players/teams that play more often. When $h = b = \gamma = 0$, the computations reproduce the Glicko ratings. The learning rate k depends on the rating's variance, allowing for faster changes when the rating is not yet well supported by evidence.

5.2.3 *Pi-ratings*

Constantinou and Fenton [27] presented a domain-specific rating system. Each team is assigned two ratings, representing its strength when playing home (R^α) and when playing away (R^β). For each match, the expected goal difference ($\widehat{\Delta G}$) is calculated based on home team's home rating (R_H^α) and away team's away rating (R_A^β).

$$\widehat{G}_H = 10^{\frac{|R_H^\alpha|}{c}} - 1 \tag{5.25}$$

$$\widehat{G}_A = 10^{\frac{|R_A^\beta|}{c}} - 1 \tag{5.26}$$

$$R_H^\alpha < 0 \implies \widehat{G}_H := -\widehat{G}_H \tag{5.27}$$

$$R_A^\beta < 0 \implies \widehat{G}_A := -\widehat{G}_A \tag{5.28}$$

$$\widehat{\Delta G} = \widehat{G}_H - \widehat{G}_A \tag{5.29}$$

where c is a metaparameter of the ratings. After a match is played, the expected goal difference is compared to the actual goal difference (ΔG), and both R^α and R^β get updated, with each of the updates having a separate learning rate.

$$e = \Delta\widehat{G} - \Delta G \tag{5.30}$$

$$\psi(e) = c\log_{10}(1 + |e|) \tag{5.31}$$

$$R_H^\alpha \mathrel{+}= \lambda\psi(e) \cdot \text{sign}(e) \tag{5.32}$$

$$R_H^\beta \mathrel{+}= \gamma\psi(e) \cdot \text{sign}(e) \tag{5.33}$$

$$R_A^\beta \mathrel{+}= \lambda\psi(e) \cdot \text{sign}(-e) \tag{5.34}$$

$$R_A^\alpha \mathrel{+}= \gamma\psi(e) \cdot \text{sign}(-e) \tag{5.35}$$

where λ and γ are the ratings' learning rates.

5.2.4 *Gaussian-OD Ratings*

Gaussian-OD ratings are an extension of the TrueSkill rating system [49]. The TrueSkill system was originally designed for ranking players in a computer game called "Halo". The motivation was to match equally skilled players against each other to maximize the overall enjoyment of the game. This further illustrates the usefulness of models presented in this thesis beyond the sole purpose of predicting future outcomes. In TrueSkill, each team is assigned a Gaussian distribution representing the user's prior about the team's skill. Unlike in Elo, the variance of each team rating is a parameter that changes value over time. In Guo et al. [52], the authors promoted a version of the TrueSkill, where each team is assigned a separate Gaussian distribution for its offensive ($p(o) := \mathcal{N}(\mu_o, \sigma_o^2)$) and defensive ($p(d) := \mathcal{N}(\mu_d, \sigma_d^2)$ skill. TrueSkill generally assumes that even if we knew the exact value of the team's skill (the variance of the

Gaussian would be equal to 0), its performance would still be stochastic, as the teams do not perform the same each day. The defensive ($p_d := \mathcal{N}(d, \beta^2)$) and offensive ($p_o := \mathcal{N}(o, \beta^2)$) performances are thus affected by the performance variance β^2. The home goals scored generation process is then assumed to be $G_H \sim \mathcal{N}(p_{o_H} - p_{d_A}, \gamma^2)$. β^2 and γ are metaparameters for performance and score variance. Finally, the prior distributions are updated after each match according to the following equations:

$$\pi_{o_H} = \frac{1}{\sigma^2_{o_H}} + \frac{1}{2\beta^2 + \gamma^2 + \sigma^2_{d_A}} \tag{5.36}$$

$$\pi_{d_A} = \frac{1}{\sigma^2_{d_A}} + \frac{1}{2\beta^2 + \gamma^2 + \sigma^2_{o_H}} \tag{5.37}$$

$$\tau_{o_H} = \frac{\mu_{o_H}}{\sigma^2_{o_H}} + \frac{G_H + \mu_{d_A}}{2\beta^2 + \gamma^2 + \sigma^2_{d_A}} \tag{5.38}$$

$$\tau_{d_A} = \frac{\mu_{d_A}}{\sigma^2_{d_A}} + \frac{\mu_{o_H} - G_H}{2\beta^2 + \gamma^2 + \sigma^2_{o_H}} \tag{5.39}$$

$$\sigma^2_{o_H} := \pi^{-1}_{o_H} \tag{5.40}$$

$$\mu_{o_H} := \sigma^2_{o_H} \tau_{o_H} \tag{5.41}$$

The equations for updating the remaining skills are analogous.

5.2.5 *Berrar Ratings*

Berrar ratings were introduced in Berrar, Lopes, and Dubitzky [13] as input features to a more complex model. The idea behind these ratings it to use a logistic function to predict the number of goals scored using once again offensive and defensive strengths of the teams. The formulas for estimating the expected goals scored are as follows:

$$\widehat{G}_H = \frac{\alpha}{1 + \exp(-\beta_H(o_H - d_A) - \gamma_H)} \tag{5.42}$$

$$\widehat{G}_A = \frac{\alpha}{1 + \exp(-\beta_A(o_A - d_H) - \gamma_A)} \tag{5.43}$$

where o and d are the offensive and defensive ratings of the competing teams, α stands for the maximum possible number of goals scored predicted. The authors set the $\alpha = 5$ as more than five goals are scored very rarely. β then determines the steepness of the logistic function, while γ defines the threshold (also known as bias). The updates to the ratings are then done in the following fashion:

$$o_H \mathrel{+}= \omega_{o_H}(G_H - \widehat{G}_H) \tag{5.44}$$

$$d_H \mathrel{+}= \omega_{d_H}(G_A - \widehat{G}_A) \tag{5.45}$$

where ω stands for the learning rate for the particular rating. Updates of the away team are done analogously.

6

VALIDATION FRAMEWORK

All the data used in this review came from the Open International Soccer Database v2 [34]. We divided the data into two sets. Matches before 07/2010 formed a validation set, used for hyperparameter tuning, and matches after 07/2010 formed a test set, used solely for evaluation. The first 5 rounds of each season were used as a burn-in period, omitted from the evaluation. This left us with 91,155 matches in the test set. The validation set was used to validate our implementations, training the parameters of subsequent regression models for rating systems, tuning hyperparameters of the models and trying out several optimization algorithms. All the presented results are computed on the test set.

The goal of the evaluation was to answer the following research questions:

1. How do the models compare in terms of predictive performance?
2. Do mathematically similar models produce similar predictions?

6.1 MODEL FITTING

The models' parameters and outputs are summarized in Table 6.1.

6.1.1 *Statistical Models*

All the statistical models are fitted by maximizing their respective weighted likelihood functions on the set of historical matches M:

$$L = \prod_{i=1}^{|M|} P\left(G_H^i = x, G_A^i = y|\theta\right) \cdot w_i. \tag{6.1}$$

where w_i represents the weight of each observation, G_H^i and G_A^i are the goals scored by home and away team in match i, and P is the probability of the respective match result as parametrized by θ. The parameters belonging to θ are summarized in Table 6.1. During the evaluation on the test set the parameters (Table 6.1) are refitted after each league's round to account for the newly obtained information. To reduce the

Table 6.1: Models' parameters and outputs overview. The τ marks parameters belonging to each team.

	Metaparameters	Parameters	Outputs
Double Poisson	α	Str_τ, H	P(HDA)
Bivariate Poisson	α	Str_τ, H, λ_c	P(HDA)
Double Weibull	α	Str_τ, H, c_H, c_A	P(HDA)
Bivariate Weibull	α	$Str_\tau, H, c_H, c_A, \kappa$	P(HDA)
Elo	k, γ, H		R_τ, E
Steph	c, h, b, γ, λ		r_τ, v_τ, e
pi-ratings	λ, γ, c		$R^\alpha_\tau, R^\beta_\tau, \widehat{\Delta G}$
Gaussian-OD	β, γ, σ_0		$\mu_{o_\tau}, \mu_{d_\tau}, \sigma_{o_\tau}, \sigma_{d_\tau}$
Berrar	$\beta, \gamma, \omega^\alpha, \omega^\beta$		$o_\tau, d_\tau, \widehat{G}_\tau$

computational time, we limit the set of historical matches M by removing matches older than 5 years in each iteration. The matches from last 5 years can be viewed as the training set for the iteration.

Since the first successful application [33], exponential time weighting is being commonly used as

$$w_i = e^{-\alpha \Delta t}, \tag{6.2}$$

where Δt is the number of days passed since the match was played and α is a metaparameter. We found $\alpha = 0.002$ to perform best on our validation set. The same value was found in Boshnakov, Kharrat, and McHale [17] and in Ley, Wiele, and Eetvelde [77] value of $\alpha = 0.0019$ was used.

6.1.2 *Rating Systems*

As the outputs of the ratings are not directly the probability distribution over the match outcomes ($P(HDA)$), a subsequent model has to be applied [57]. We use multinomial logistic regression for this purpose. The parameters of the regression are optimized inside the meta-optimization routine for finding the rating's metaparameters. First, the ratings' computations (given the current set of metaparameters) are carried out through the data. The pre-match ratings then serve as input to the regression model. The regression model then produces in-sample predictions based on the given features. The in-sample loss is then reported as the loss belonging to the current set of metaparameters. The meta-optimizer then selects another set of metaparameters to be evaluated, and the process is repeated. The inner routine is summarized in the following pseudo-code:

```
def optimize_rating(data, results, metaparams, loss_func)
    ratings = compute_ratings(data, results, metaparams)
    lr = LogisticRegression()
    lr = lr.fit(ratings, results)
    predictions = lr.predict_proba(ratings)
    loss = loss_func(predictions, results)
    return loss.mean()
```

We observed that multiple runs of meta-optimization result in different metaparameters while achieving the same loss. We therefore do not state any concrete values of the metaparameters found. While the parameters of the regressors and the metaparameters (Table 6.1) of the rating systems are determined on the validation set, the ratings (denoted as "Outputs" in Table 6.1) are updated after each match in the test set to serve as input features for the regressor in the next league round.

The optimization of the regression parameters has been done by the L-BFGS-B algorithm [22]. Optimization of the ratings' hyperparameters has been carried out by the PSO [64], as was done in Berrar, Lopes, and Dubitzky [13]. We experimented with other meta-optimization techniques but they were inferior to the PSO in terms of predictive performace and (mainly) computational time.

During the validation, we noticed, that the meta-optimization of the Berrar ratings fails to converge occasionally even when given more iterations. We have resolved the issue by halving the number of metaparameters. In the original model the metaparameters for updating the ratings are separate for home and away team. As stated in Table 6.1, we use the same metaparameters for the home and away team ratings updates.

6.2 EVALUATION MEASURES

Besides crossentropy (Eq. 2.6), we evaluate the models using Ranked Probability Score and Accuracy. Moreover, to quantify the similarities between the models' predictions we use the Jensen-Shannon divergence.

6.2.1 *Ranked Probability Score*

The ranked probability score was proposed by Epstein [36] for evaluating ordinal outcomes. For the ternary outcome game of soccer, the formula is as follows:

$$\mathrm{RPS}(\mathbf{p},\mathbf{y}) = \frac{1}{2}\sum_{i=1}^{2}\left(\sum_{j=1}^{i}(p_j - y_j)\right)^2, \tag{6.3}$$

where p_j is the estimated probability of outcome j, and $y_j \in \{0,1\}$, with $y_j = 1$ indicating that outcome j was realized. The suitability of using this metric for evaluating soccer outcome predictions was heavily proclaimed in Constantinou and Fenton [26] and has been widely used ever since.

6.2.2 *Accuracy*

Accuracy serves as the most crude evaluation measure. It simply represents how many times on average the outcome with the highest estimated probability was realized.

6.2.3 *Similarity Measures*

Besides the predictive performance of the models, we are also interested in analyzing how much the predictions of the models differ from each other since there are many similarities both among the statistical models and among the ratings. For this purpose, we compute the average Jensen-Shannon divergence between the models' predictions. The Jensen-Shannon divergence between two probability distributions P and Q is given by:

$$JSD(P \mid\mid Q) = \frac{1}{2} D_{KL}(P \mid\mid M) + \frac{1}{2} D_{KL}(Q \mid\mid M) \tag{6.4}$$

$$M = \frac{1}{2}(P + Q) \tag{6.5}$$

$$D_{KL}(P \mid\mid Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \tag{6.6}$$

7

RESULTS

The results are summarized in Table 7.1. The first thing to notice is that the models' performances are very close to each other. The Berrar ratings achieved best results by both the RPS and xEnt measures, while the Double Weibull and Bivariate Weibull models reached the highest accuracy score. The Double Weibull model placing ahead of its Bivariate variant might look suspicious at first. However, during the validation, we noticed that while the Bivariate Weibull sometimes provided the best fit for the training data, the performance did not always translate into the test set. This suggests that finding the right dependence parameter κ is difficult. It is remarkable how well the general rating system Elo with only minor modifications works for soccer. This is not the case for the other two general rating systems – the Steph ratings and the Gaussian-OD ratings. Another result that catches the eye is the performance of the Double Poisson model. This only confirms its competitiveness, as suggested by Ley, Wiele, and Eetvelde [77]. The only model that significantly falls behind are the Gaussian-OD ratings.

Table 7.1: Comparison of the tested models via the evaluation metrics.

	xEnt	RPS	Acc.
Berrar	1.0246	0.2101	48.54
Bivariate Poisson	1.0251	0.2103	48.58
Double Poisson	1.0254	0.2103	48.57
Double Weibull	1.0255	0.2103	48.60
pi-ratings	1.0258	0.2103	48.56
Bivariate Weibull	1.0260	0.2105	48.60
Elo	1.0263	0.2105	48.49
Steph	1.0291	0.2114	48.26
Gaussian-OD	1.0347	0.2134	47.84

Table 7.2: Average Jensen-Shannon divergence between the models' predictions. BP = Bivariate Poisson, BW = Bivariate Weibull, DP = Double Poisson, DW = Double Weibull.

	Berrar	BP	BW	Elo	DP	DW	pi-rtgs	Steph	Gauss
Berrar	0.000	0.051	0.054	0.027	0.051	0.051	0.030	0.040	0.063
BP	0.051	0.000	0.011	0.052	0.014	0.014	0.052	0.058	0.071
BW	0.054	0.011	0.000	0.054	0.023	0.017	0.054	0.061	0.075
Elo	0.027	0.052	0.054	0.000	0.051	0.051	0.024	0.032	0.062
DP	0.051	0.014	0.023	0.051	0.000	0.009	0.051	0.056	0.069
DW	0.051	0.014	0.017	0.051	0.009	0.000	0.050	0.057	0.070
pi-rtgs	0.030	0.052	0.054	0.024	0.051	0.050	0.000	0.036	0.064
Steph	0.040	0.058	0.061	0.032	0.056	0.057	0.036	0.000	0.058
Gauss	0.063	0.071	0.075	0.062	0.069	0.070	0.064	0.058	0.000

7.1 PREDICTIONS' SIMILARITY

As the evaluation metrics of plurality of the models are very close to each other, it is also not surprising that the actual predictions exhibit many similarities, too, as shown in Table 7.2. Even without knowing that there are two classes of models, we would be able to distinguish the statistical models from the ratings based purely on their prediction similarities. We can observe that especially the predictions of the statistical models are very close to each other. This behavior was anticipated, as the very definitions of the models are very similar (with a certain parameter setup, they all reduce to the Double Poisson model). The similarity of the Gaussian-OD to other models is lower mostly due to its inferior performance.

7.2 MODEL ADAPTABILITY

Another property of the models we were interested in is how quickly they adapt to new information. Between the seasons, the team's composition, and therefore also its strength, can change dramatically. We thus divided the matches into 10 groups based on which part of the season they occurred (Figure 7.1).

The plot shows that the models' performance is generally higher in the second half of the season when more data are available. The statistical models trail behind the rating systems in the first third of the season, while providing a generally better fit in the second half of season. Berrar rating seems to outperform the competition in most parts of the season. The slight decrease in the models' performances, right after half of the season was played, might be due to breaks in the schedule that typically occur in

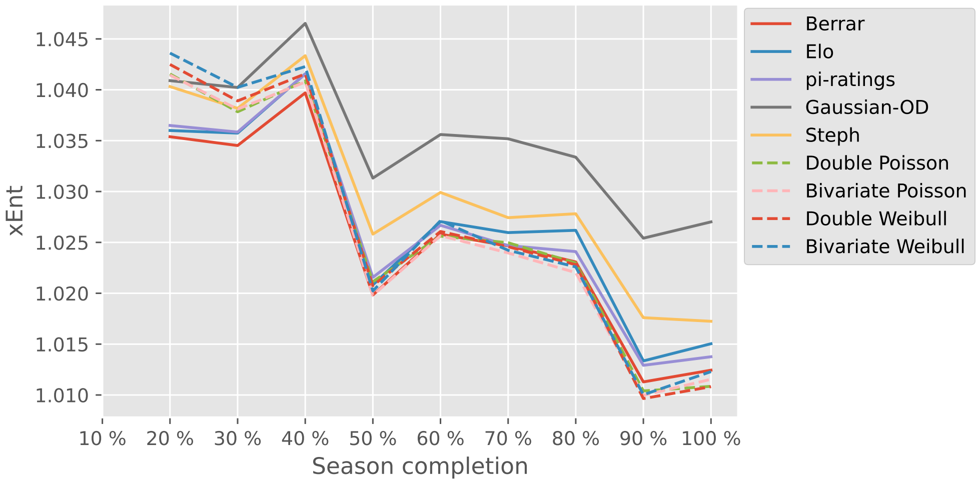

Figure 7.1: Error of the models, as measured using the crossentropy (Eq. 2.6), as a function of percentage of season completed.

the middle of a season. Another explanation could be that in some leagues, there are transfer windows opened during this period of time, which could lead to changes in teams' compositions.

8

CONCLUSION

We reimplemented a wide selection of the top-performing score-based models for soccer outcome forecasting from the past decades and benchmarked them on the largest soccer dataset published to date. We asked two core research questions regarding the models' performances and similarities. We conclude from the experiments that the individual predictions, as well as the overall performances, were very similar across the top models tested, likely suggesting the limits of this generic approach to score-based match outcome modeling. Additionally, we observed that the rating systems adapt faster to changes in teams' strengths and achieve better performance at the beginnings of the seasons, while the statistical models catch up and take a small lead at the ends.

Our results suggest that any dramatic improvement in the predictive performance of any rating or statistical method seems unlikely now. We therefore propose that further research should attempt to address the problem with a significantly different "class" of models (Part iii). These could possibly produce more diverse predictions, opening new possibilities for ensembling and other machine learning techniques.

Part III

PROPOSED MACHINE LEARNING MODELS

9

INTRODUCTION

In this part of the thesis, we propose our own models to tackle the problem of score-based forecasting. As we have seen in our experimental review (Part ii), there are several existing models with very similar performance. While there have been attempts to predict the winners of a match using machine learning, these approaches usually rely on including other features than those that can be directly derived from the final scores [46, 51, 56] and thus cannot be compared with models introduced in Part ii. Purely score-based approaches were summarized in Section 4.3. We aim to close this gap and leverage a machine learning approach based on careful feature engineering.

Moreover, as the data we are dealing with are relational in nature, we propose a relational version of an embedding model. We argue that incorporating relational learning techniques might benefit the field considerably. It only seems natural as the data arising from sports records possess interesting relational characteristics on many levels of abstraction, from the matches themselves forming relations between teams, players and seasons, to the course of the individual matches being driven by the rules of each sport with game-play patterns stemming from these. We propose simple relational representations, background knowledge, and modeling concepts for which we provide some interpretable insights. Particularly, we focus on expressing a concept we called "Lifted relational team embeddings" in the framework of Lifted relational neural networks (LRNNs) [118], combining relational fuzzy logic with gradient descent optimization.

The rest of this part is organized as follows. In Section 10 we describe the types of predictive models considered. Section 11 describes the features we constructed for the feature-based models. In Section 12 we validate the different modeling approaches on the disclosed data set. Section 13 provides a discussion of the principal trends observed in the experimental results.

10

MODELS

In what follows, the terms *loss* and *win* associated with a match refer to the home team's outcome in that match, unless stated otherwise. Here we discuss the methods we used to estimate

$$(p_l, p_d, p_w) \in [0,1]^3, \text{ such that } p_l + p_d + p_w = 1 \tag{10.1}$$

i.e., the probabilities of the three possible outcomes *loss*, *draw*, *win* of a given match.

10.1 BASELINE PREDICTORS

We introduce two reference prediction policies, intended to act as natural *upper* and *lower bounds* on the prediction errors achievable with the trainable models introduced later.

The naive policy corresponding to the upper-bound predicts (10.1) for each match in a given season and league by setting p_l to be the proportion of home-team losses in all matches of that league in the previous season, and similarly for p_d and p_w. Intuitively, this predictor exploits the usual home-team advantage [106], which is quantified into the probabilities using the relative frequencies from the immediately preceding season. Failing to improve on such a prediction policy would indicate a useless predictor. The likely lower bound on prediction error is provided by bookmaker's data. Bookmakers are considered a very reliable source of predictions [40]. Bookmaker's odds represent inverted probability estimates of the outcomes. However, to get an edge over the market, the bookmaker employs a so-called margin, resulting in the inverted probabilities summing up to more than 1. Therefore we normalized the probability triple with a common divisor to make it sum up to one. For example if the odds for the home team to win, draw and lose are (respectively) 1.89, 3.13, 5, the implied inverted probabilities are 1.89^{-1}, 3.13^{-1}, 5^{-1}, and the normalized probabilities are $1.89^{-1}/Z$, $3.13^{-1}/Z$, $5^{-1}/Z$ where $Z = 1.89^{-1} + 3.13^{-1} + 5^{-1}$. More advanced methods for deriving probabilities from the odds are described by [122]. Improving on such a predictor is unlikely as the bookmakers have access to much more detailed data.

10.2 FEATURE-BASED CLASSIFICATION MODEL

While the raw data [34] do not contain any features besides the competing teams' names, numerous features can be derived from such data using feature engineering techniques (Section 11). Once we transform the data into a tabular dataset, we can choose from a plethora of conventional feature-based machine learning models. We naturally select from multi-class classifiers yielding a probability distribution on target classes.

From among such eligible classifier types, we chose *Gradient Boosted Trees* [43]. This choice was motivated by a multitude of machine learning competitions such as those hosted by Kaggle[1], where the Gradient Boosted Trees algorithm, and specifically its Xgboost implementation [23], turns out to be highly successful for problems of a similar character.

10.3 FEATURE-BASED REGRESSION MODEL

The loss function minimized by Xgboost during classifier fitting is the crossentropy (Eq. 2.6). This loss function does not reflect the intuitive order

$$loss < draw < win \tag{10.2}$$

on classes. However, the Xgboost algorithm can also be run in a *regression mode*, where the resulting model yields real numbers. We leveraged this mode to accommodate the order (10.2) by representing the three classes as $0, 0.5, 1$, respectively. In the regression setting, the standard squared loss is minimized through training.

To map a model's output $r \in [0;1]$ to the required distribution (10.1), we introduce an additional trainable model component. Specifically, we posit for each $i \in \{l, d, w\}$ that

$$p_i(r) = \frac{f_i(r)}{f_l(r) + f_d(r) + f_w(r)} \tag{10.3}$$

where $f_i(r)$ is modeled as a beta distribution

$$f_i(r) = \mathrm{Beta}_{\vec{\Theta}}(r) \tag{10.4}$$

in which the parameters $\vec{\Theta}$ maximize the function's fit with tuples $r, P_i(r)$ available in training data; in particular, $P_i(r)$ is the proportion of training examples with outcome i among all examples for which the model yields r. Intuitively, e.g. $f_w(r)$ is an estimate of the win-probability for a match with regressor's output r. Note that in general $f_l(r) + f_d(r) + f_w(r) \neq 1$, hence the normalization in (10.3).

[1] A platform for hosting machine learning competitions at https://kaggle.com

10.4 LIFTED RELATIONAL NEURAL NETWORKS

LRNNs [118] is a relational learning framework utilizing a parametrized fragment of relational fuzzy logic as a language for representation of various models and a gradient descend technique for their parameter training. The model representation can be viewed as a lifted *template* for neural networks, as it enables neural computations to be performed upon relational data by constructing a different computational graph, or neural network, for each of the differently structured relational examples. Generalization onto unseen relational examples is due to the joint derivation of structure and parameters of these neural networks from the single lifted template.

For a regular training of an LRNN, as we do in experiments reported in this thesis, one firstly needs to manually create the template, which may encode some background knowledge, or intuition, together with various modeling constructs. Secondly, one needs learning examples encoded in relational logic together with corresponding target predicate labels. Subsequently in the learning process, the LRNN engine grounds the template w.r.t. the different examples to create the corresponding neural networks, which are then jointly trained w.r.t. the labels, in a manner similar to that of standard deep learning frameworks.

10.4.1 *Knowledge Representation*

In its raw form, the match records contain merely the team names and the result. For the LRNNs we had to derive appropriate relational representation. Since they learn from Herbrand interpretations, we encoded the records with numerical outcomes into predicates, which we describe in Table 10.1.

It can be noted that these predicates encode the match records in a very straightforward manner. We also incorporated some derived predicates with simple domain knowledge, such as the goal difference and recency, which we generally expected to play some role in the predictions.

10.4.2 *Lifted Relational Team Embeddings*

Here we describe the proposed relational embedding model as expressed in the language of LRNNs. Firstly, we tested the hypothesis that there exists some predictive latent space embedding the teams. This is based on an intuition from various rating systems (Section 5.2), where each team is assigned one or more parameters denoting its particular strength, possibly within different areas, such as when playing at home stadium and when playing away. However, opposite to the existing rating systems, the idea of the embedding approach is to explore meaning of these latent parameters automatically by the means of regular learning from data. In other words we do not explicitly predefine what particular types of strengths we are looking for, and rather let

Table 10.1: Overview of predicates extracted from the data for the relational learners.

Predicate	Description
home(Tid)	Team Tid is home team w.r.t. prediction match.
away(Tid)	Team Tid is away team w.r.t. prediction match.
team$(Tid, name)$	Team Tid has name $name$.
win(Mid, Tid_1, Tid_2)	Win of home team Tid_1 over away team Tid_2 in match Mid.
draw(Mid, Tid_1, Tid_2)	Draw between home team Tid_1 and Tid_2 in match Mid.
loss(Mid, Tid_1, Tid_2)	Loss of home team Tid_1 to team Tid_2 in match Mid.
scored(Mid, Tid, n)	The team Tid scored more than n goals in match Mid.
conceded(Mid, Tid, n)	The team Tid conceded more than n goals in match Mid.
goal_diff(Mid, n)	Difference in goals scored by the teams is greater than n.
recency(Mid, n)	The match Mid was played more than n rounds ago (w.r.t. prediction match).

the most predictive types be explored directly from data w.r.t. to given learning target. We can encode this scenario in LRNNs as follows.

$w_1^{(0)}$: *type1*$(T) \leftarrow$ *team(T, chelsea)*

$w_2^{(0)}$: *type1*$(T) \leftarrow$ *team(T, arsenal)*

...

$w_i^{(0)}$: *type2*$(T) \leftarrow$ *team(T, chelsea)*

...

$w_j^{(0)}$: *type3*$(T) \leftarrow$ *team(T, everton)*

where the types $type_1 \ldots type_3$ denote individual embedding dimensions of the teams.

The introduced modeling concept is directly based on the idea of *soft clustering* from [118], where the goal is to explore latent predictive types of domain elements. We may directly use aggregation of such embeddings for prediction of outcome of *home* vs. *away* team matches using the following rules.

$$
\begin{aligned}
&w^{(1)}_{(1;1)} : outcome \leftarrow home(T1) \wedge type1(T1) \wedge away(T2) \wedge type1(T2)\\
&w^{(1)}_{(1;2)} : outcome \leftarrow home(T1) \wedge type1(T1) \wedge away(T2) \wedge type2(T2)\\
&\dots\\
&w^{(1)}_{(3;3)} : outcome \leftarrow home(T1) \wedge type3(T1) \wedge away(T2) \wedge type3(T2)
\end{aligned}
$$

This construct in principle creates a fully connected neural network with one hidden embedding layer, such as e.g. in the famous *word2vec* embedding architecture [91]. For all the historical matches we then jointly perform corresponding gradient updates of the weights to reflect the actual values of the *outcome* labels. We further denote this architecture as *embeddings*.

In theory, the embeddings possibly capture some information on the relational interplay between the matches as they are jointly optimized on the whole match history. However, we find this approach quite limited as it is rather naive to expect the flat, fixed-size embeddings to reflect all the possible nuances of the complex relational structure stemming from the different outcomes of different historical matches played between different teams in different orders. Moreover, the embedding space dimensions are fixed while all the possible relational histories are obviously not bounded in that way. On the other hand, despite often disregarding the relational information, embeddings have experimentally proved quite strong in exploiting the relevant features, even from data considered relational [16]. Fortunately with LRNNs, we can easily capture the relational structures explicitly while keeping the benefits of embedding learning. For that we first extend the template with a predicate capturing the different outcomes of *historical* matches (w.r.t. prediction match) through a learnable transformation as

$$
\begin{aligned}
w^{(2)}_1 : &\quad outcome(M, H, A) &\leftarrow &\quad win(M, H, A)\\
w^{(2)}_2 : &\quad outcome(M, H, A) &\leftarrow &\quad draw(M, H, A)\\
w^{(2)}_3 : &\quad outcome(M, H, A) &\leftarrow &\quad loss(M, H, A)
\end{aligned}
$$

with which we accordingly extend the predictive rules as

$$\begin{aligned}
w^{(1)}_{h-h(1;1)} &: outcome \leftarrow home(T1) \wedge type1(T1) \wedge outcome(M, T1, T2) \\
&\qquad\qquad\qquad\qquad \wedge type1(T2). \\
w^{(1)}_{h-a(1;1)} &: outcome \leftarrow home(T1) \wedge type1(T1) \wedge outcome(M, T2, T1) \\
&\qquad\qquad\qquad\qquad \wedge type1(T2). \\
w^{(1)}_{h-h(1;2)} &: outcome \leftarrow home(T1) \wedge type1(T1) \wedge outcome(M, T1, T2) \\
&\qquad\qquad\qquad\qquad \wedge type2(T2). \\
&\qquad\qquad \dots \\
w^{(1)}_{a-a(3;3)} &: outcome \leftarrow away(T1) \wedge type3(T1) \wedge outcome(M, T2, T1) \\
&\qquad\qquad\qquad\qquad \wedge type3(T2).
\end{aligned}$$

reflecting the possible settings of *historical* home and away positions of the *actual* home and away teams in all historical matches played. By grounding this template, the LRNN engine assures to create the corresponding relational histories transformed into respective, differently structured, neural networks. We denote this architecture as *relational embeddings*. These embeddings of teams extracted from the model learned to predict home team win can be seen in Figure 10.1.

Similarly to the other learners, these rules can be further extended by adding more contextual information on the individual matches, e.g. the goal difference.

$$\begin{aligned}
&w^{(4)}_1 : goal\_diff(M) \leftarrow goal\_diff(M, \text{-}3) \\
&w^{(4)}_2 : goal\_diff(M) \leftarrow goal\_diff(M, \text{-}2) \\
&\dots \\
&w^{(4)}_7 : goal\_diff(M) \leftarrow goal\_diff(M, 3)
\end{aligned}$$

Also recency of the match and other derived features may be incorporated, however in our preliminary experiments we did not find any significant improvements from these extra features.

10.5 MODEL PORTFOLIOS

Here we address the heterogeneity among different soccer leagues by exploring *model portfolios*. Briefly, the set of all considered leagues is first split into relatively homogeneous partitions and a model is learned for each partition separately. The portfolio then collects all these models and for each prediction, it invokes only the applicable model.

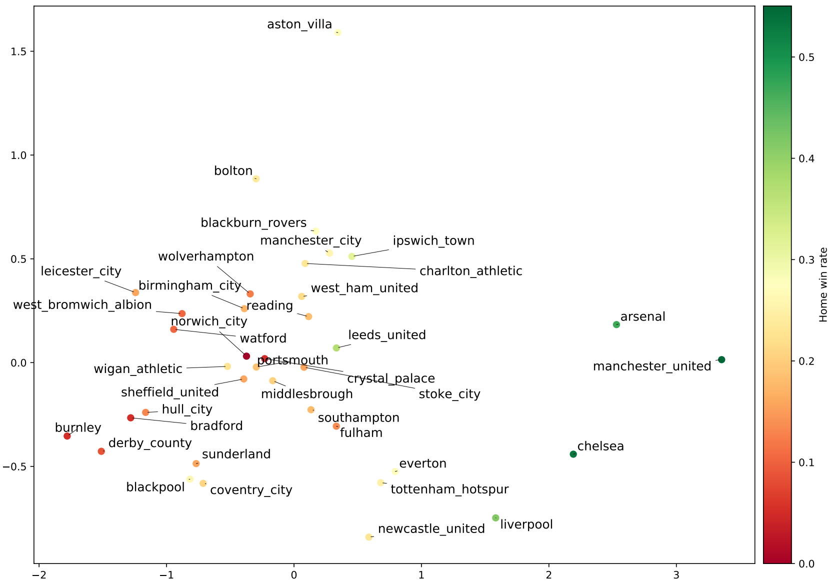

Figure 10.1: Visualization of PCA projection of the learned embeddings of individual teams from the home-win model. A significant relationship between the home win rate, captured by the colorscale, and the variance captured by the main X axis can be observed.

The type of the constituting models can be any of those described in the preceding sections.

The mentioned heterogeneity is due to several aspects. First of all, each league has a different structure. Most often each two teams from the same league play against each other two times in one season, once at each respective home stadium. But there are also leagues where two teams meet only once or where the teams are divided into groups and a team plays only other teams within the same group. Moreover, the number of teams that are promoted and relegated is also different for each league.

Besides the structural differences, the leagues differ in play-style. Some leagues favor offensive style while others play more defensive soccer, resulting in discrepancies in statistics like draw percentage, average number of scored goals (the latter shown in Figure 10.2), etc. Moreover, the number of given yellow and red cards per match varies between leagues, possibly leading to larger changes in the team strengths between consecutive rounds, resulting from the offender being disqualified from one or more consecutive matches.

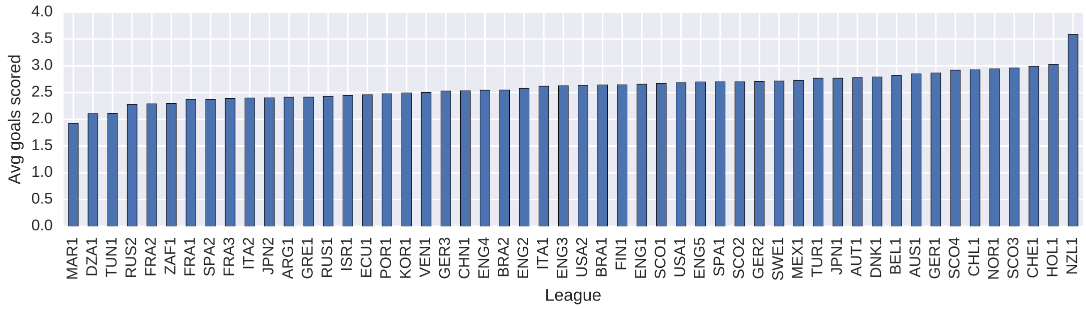

Figure 10.2: Overall average number of goals scored per match in different leagues.

We tried three different ways to partition the league set. The first was to cluster the leagues according to the performance of a selected "ethalon" predictor, dissecting the list of leagues sorted by this indicator into the tough group, the easy group, and so on, depending on the chosen granularity. The second approach was to cluster the leagues based on the leagues' features (Table 11.2) using the standard k-means algorithm with the Euclidean distance on normalized numeric features. The last method was to train a separate model for each league.

11

FEATURE ENGINEERING

For the predictors in Sections 10.2 and 10.3, we need a set of relevant features for each learning sample corresponding to a match. The set of features we constructed is listed in Table 11.1 and we describe their categories in turn. With exceptions indicated in the table, each feature relates to a team and so appears twice in the tuple describing a match, once for each of the two teams. The features are not evaluated for samples in the first two seasons due to the time lag required for some of them.

11.1 HISTORICAL STRENGTH

To reflect the long-term strength of the teams, we extracted means and variances of the scored and conceded goals, win percentages, and draw percentages. These statistics are calculated separately for matches played home and away as a team playing home is typically stronger than when playing away [106].

The statistics are aggregated from the current and the two preceding seasons.

11.2 CURRENT FORM

Even the strongest teams can have a period of weaker performance during a season and vice versa. Therefore we also include in the feature set a set of statistics similar to the above, except aggregated only over the last five matches played by the concerned team. If less than five matches have been played by the team in the current season, the feature is not evaluated and acquires a missing-value indicator. These statistics are not computed from home and away games separately as such split statistics would aggregate a very small number (2 or 3) matches.

Additionally, the current strength of the team could be affected by the number of days since last match because of fatigue. Therefore the number of rest days is also included as a feature.

Table 11.1: Summary of constructed features. Except features shown in *italics* each feature appears twice in the description of a sample; referring respectively to the home team and the away team. Moreover, features prefixed H/A are computed separately from team's home and away matches respectively.

Historical strength	computed from matches from the current and last 2 seasons
H/A WIN PCT	winning percentage
H/A DRAW PCT	drawing percentage
H/A GS AVG	goals scored average
H/A GC AVG	goals conceded average
H/A GS STD	goals scored standard deviation
H/A GC STD	goals conceded standard deviation
Current Form	computed from the last 5 matches played
WIN PCT	winning percentage
DRAW PCT	drawing percentage
GS AVG	goals scored average
GC AVG	goals conceded average
GS STD	goals scored standard deviation
GC STD	goals conceded standard deviation
REST	number of days since team's last match
Pi-ratings	computed from matches from the current and last 2 seasons
H/A RTG	pi-rating
EGD	expected goal difference by pi-ratings
PageRank	computed from matches from the current and last 2 seasons
EPTS PR	PageRank computed from graph weighted by expected points
Match importance	
T↑	relative point differences between the team and the teams on first 5 positions in league table
T↓	relative point differences between the team and the teams on last 5 positions in league table
RND	league round

11.3 pi-ratings

These features relate to the pi-ratings introduced in Section 5.2.3. We included directly the home and away ratings of each of the two teams and the predicted goal difference between the two.

11.4 pagerank

A drawback of the historical strength features is that they do not account for the opposing teams' strengths in historical matches. A decisive win against a weaker opponent might not be as important as a close win against a title contender. To account for this factor, we utilized the PageRank [72] algorithm. PageRank was originally developed for assessing the importance of a website by examining the importance of other websites referring to it. Similarly, our assumption was that a strong team would be determined by having better results against other strong teams.

The PageRank of a team can be computed out of a matrix with columns as well as rows corresponding to teams. Each cell holds a number expressing the relative dominance of one team over the other in terms of previous match outcomes. In particular, the i, j cell contains

$$\frac{3w_{ij} + d_{ij}}{g_{ij}}, \tag{11.1}$$

where w_{ij} (d_{ij}) is the number of wins (draws) of team i over (with) team j, and the normalizer g_{ij} is the number of games played involving the two teams. These numbers are extracted from the current and the two preceding seasons. The coefficient 3 reflects the standard soccer point assignment.

In comparison with pi-ratings, PageRank does not work with the actual goal difference but solely with the match outcomes. The pi-ratings can experience larger changes after a single round, while the PageRank is calculated just form a slightly modified matrix. We thus consider PageRank a more regularized counterpart of pi-ratings.

11.5 match importance

Match importance can be reasonably expected to affect players' performance and so represents a relevant feature. It is however not obvious how to estimate it.

Match importance is closely tied with team's rank and current league round. Adding the league round number to the feature vector is thus straightforward. However, dealing with team's rank is more complicated. First of all, the ranking of teams with the same number of points is calculated by different rules in each league. More importantly, the ranking is often too crude to capture the match importance, because it neglects the point differences. For instance in a balanced league, a team can be in 5^{th} place, trailing by only few points to the team in first place, with several rounds to go in the season, while in a league dominated by few teams, a team in 5^{th} position would have no chance in the title race, reducing the importance of the remaining games. There were attempts to model the match importance by simulating the remaining matches of a season [70]. However, a quantity that can only follow from computational simulations can hardly be expected to affect the player's mindsets.

We decided to extract the points from the league table, from which we subtracted the points of the team in question, obtaining relative point differences. The points are accumulated as the season goes on, and normalized by the number of games played so far. The relative point differences for team i were aggregated in vector $T_i(k)$ such that

$$T_i(k) = \frac{\pi_{\mathsf{rank}(k)} - \pi_i}{g_i}, \tag{11.2}$$

where k ranges from 1 to the number of all teams in the league, π_i ($\pi_{\mathsf{rank}(k)}$) is the number of points team i (team ranked k-th, respectively) accumulated through the season, and g_i is the number of games team i played.

To extend the feature set with a fixed number of scalars, we extracted only the first five and last five components of T_i corresponding to the head and the tail of the ranking.

11.6 LEAGUE CHARACTERISTICS

League-specific features consist of the numbers of teams, rounds, home win percentages, draw percentages, goal difference deviations, and home/away goals scored averages. These statistics are meant to provide a context for the historical strength features introduced earlier. For instance, scoring 2.5 goals per match on average has a different weight in a league where the average is 2 goals per match, and one with the average of 3 goals per match.

Table 11.2: Features used for league clustering as well as inputs to the models. H/A stands for home/away.

H/A GS AVG	goals scored average in last 2 seasons
H/A GS STD	goals scored standard dev. in last 2 seasons
H/A WIN PCT	winning pct. in last 2 seasons
DRAW PCT	drawing percentage in last 2 seasons
TEAM CNT	number of teams in last season
GD STD	standard dev. of goal difference in last 2 seasons
RND CNT	number of rounds played in last season

12

EXPERIMENTAL EVALUATION

Here we provide details on our evaluation framework. The models introduced in Section 10 we designed for a Soccer Prediction Challenge 2017 organized by the Machine Learning journal [12]. The challenge required the participants to train their models with all data available and submit their predictions for the upcoming matches (the test set). As the upcoming matches were known (unlike the results obviously) we used only the data from the leagues, that a appeared in the test set (Table 12.1). However, to compare fairly with the state-of-the art (Section 7) we had to reffit the model in compliance with the validation framework introduced in Section 6. The following sections describe how we chosed the final model for the prediction challenge. Section 12.7 compares the selected model with the state-of-the-art.

12.1 VALIDATION AND PARAMETER TUNING

Both the relational method (Section 10.4.2) and the feature-based methods (Sections 10.2 and 10.3) require to set a few hyper-parameters. In particular, the pi-ratings learning rates ($\lambda = 0.06, \gamma = 0.5$) need to be determined. The feature-based methods additionally require setting for the Xgboost's parameters *max_depth* (= 4), *subsample* (= 0.8), *min_child_weight* (= 5) and *colsample_bytree* (= 0.25). For LRNNs we set the learning rate (0.1) and number of learning steps (50). The number of trees for Xgboost was determined using internal validation with early stopping. The rest of the parameters were tuned exhaustively through grid search, by training on the same data split, and validating on the remaining data. Ranges of values tried in the grid search were following: $\{3, 4, \ldots, 8\}$ for *max_depth*, $\{0.5, 0.6, \ldots, 1\}$ for *subsample*, $\{3, 4, ..., 8\}$ for *min_child_weight*, and $\{0.2, 0.25, ..., 0.5\}$ for *colsample_bytree*. For each of the models, we picked the parameters minimizing the RPS on the validation set.

With the exception on Section 12.2, which follows a time-wise evaluation, all the reported RPS are averages over seasons 2010/11 and further. For each season from this testing period, the model was trained on all preceding seasons.

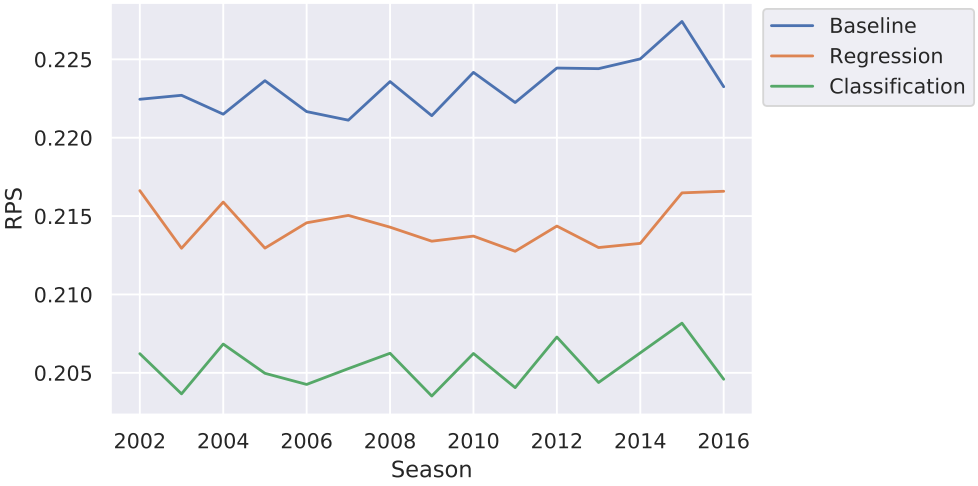

Figure 12.1: RPS of different types of predictive models over the course of several seasons (lower is better). Note the restricted scale on the vertical axis.

12.2 MODEL PERFORMANCE IN TIME

Figure 12.1 shows the RPS values for successive seasons from the third season on, so that historical strength features can be calculated from a sufficient history. The RPS is calculated only on the leagues known to be included in the challenge test set.

The two types of feature-based model types (regression, classification, c.f. Sections 10.2-10.3) as well as the upper-bound baseline (Section 10.1) are shown. The lower-bound baseline is not included as the bookmaker's odds data are not available for all leagues; we shall compare to this baseline separately.

Each RPS value in the diagram pertains to the prediction made by a model trained on all data up to (and excluding) the current season, i.e. models are retrained at every season's beginning.

A remark is in order regarding the training of the regression model. As explained earlier, besides fitting the regressor itself, we also need to train the mapping from its output to the predicted distribution. For the latter, as an over-fitting prevention measure, the proportions $P_i(r)$ (c.f. Section 10.3) are calculated as follows. When training the model for the n-th season, the r's and the corresponding proportions $P_i(r)$ are collected from all of the preceding seasons; for each k-th ($k < n$) season, they are obtained with the model learned for that season (i.e., on data from seasons 1 to $k-1$), making predictions on the k-th season. This way, the proportions following from model predictions are collected from data not used for training the models.

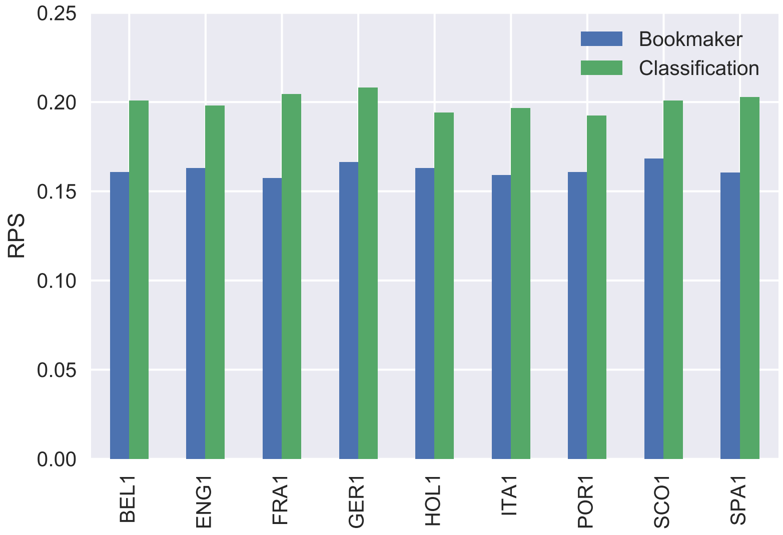

Figure 12.2: RPS comparison of the feature-based classification model with the bookmaker's predictions.

12.3 COMPARISON TO BOOKMAKER'S PREDICTIONS

For leagues where bookmaker's odds are available, we compared the best performing model (i.e. the feature-based classification model) with a predictor implicitly defined be these odds as described in Section 10.1. We downloaded odds[1] for more than 22000 matches for the 2008-2015 period.

Figure 12.2 shows the average RPS for the two predictors on individual leagues. Here, the classification model is trained as in Section 12.2, i.e. on all data preceding the season where prediction takes place.

It follows that the bookmaker completely dominates the learned classifier's predictions. That is no surprise, given the additional sources of information available to the bookmaker. These include detailed play statistics collected from the matches, changes in teams' rosters as well as video footages of the matches.

12.4 MODEL PORTFOLIO PERFORMANCE

We assessed the potential of the portfolio strategy as described in Section 10.5. We first partitioned leagues according to predictability by a model. In particular, the leagues were ranked by the RPS achieved by the feature-based classification model validated in

[1] Odds from bet365 available at http://www.football-data.co.uk/ were used.

Table 12.1: Clusters obtained by clustering the leagues by performance and by feature similarity.

	by RPS
1	AUT1, CHN1, ENG1, GRE1, HOL1, POR1, TUN1
2	BEL1, CHE1, ITA1, MAR1, SCO1, SPA1, VEN1
3	FRA1, FRA2, GER1, GER2, ISR1, KOR1, RUS1
4	CHL1, ECU1, ENG2, JPN1, MEX1, USA1, ZAF1
	by league features
1	CHN1, ENG1, FRA1, GRE1, MAR1, RUS1, TUN1
2	CHE1, ISR1, JPN1, KOR1, POR1, SCO1, SPA1, ZAF1
3	AUT1, BEL1, GER1, GER2, HOL1
4	CHL1, ECU1, ENG2, FRA2, ITA1, MEX1, USA1, VEN1

Table 12.2: RPS of a portfolio model with different league partitionings.

method	RPS
no split	0.2055
split by similarity	0.2063
split by performance	0.2064
split by league	0.2081

seasons 2007/08 – 2009/10 and trained on the preceding seasons. Then we split the list into 4 groups of 7 teams successive in this ranking. Next we produced an alternative partitioning by the *League* features (Table 11.1) through the standard k-means algorithm, setting $k = 4$. We run the stochastic k-means algorithm several times and used the clustering consisting of most equally sized clusters. Lastly, we produced singleton clusters, one for each league. The groupings achieved by the former two approaches are summarized in Table 12.1.

We trained the portfolio model using the feature-based classifier as the constituting model type. Table 12.2 presents the RPS for the three clustering variants, with models trained on seasons up to and including 2009/10 and validated on all the subsequent seasons. The results indicate a detrimental effect of each clustering variant, likely following from the smaller training sets available for training each constituting model.

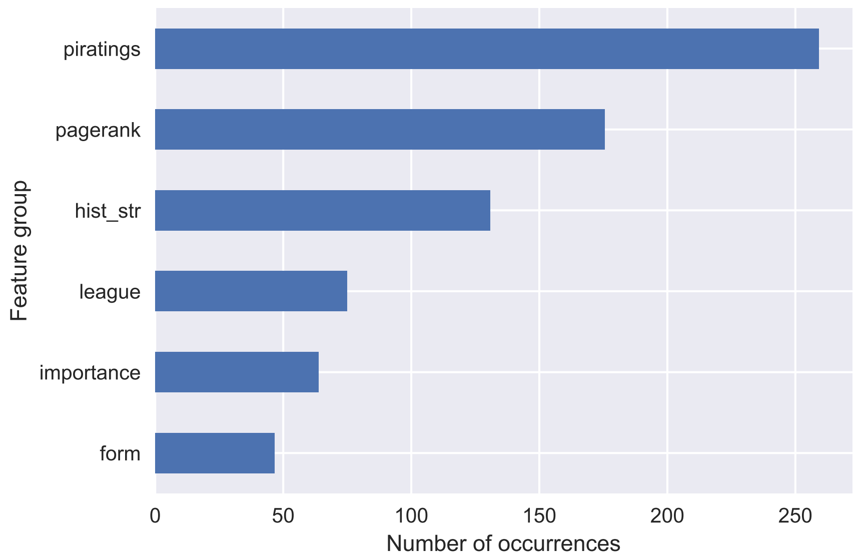

Figure 12.3: Average occurrence counts of features of given categories in the classification trees' nodes.

Table 12.3: Performance of the classification model trained on different subsets of features.

Features	RPS
pi-ratings only	0.2067
pi-ratings + PageRank + historical strength	0.2061
all feature categories	0.2055

12.5 Feature Importance

Lastly, we examined the effect of individual feature categories as defined in Section 11. We did this in two manners.

Firstly, we counted how many times a feature was used in the tree nodes of the classification model. Figure 12.3 shows that the pi-ratings were by far the most commonly used features. On the other hand, current-form features were used only sporadically.

Secondly, we trained the model (again on seasons up to and including 2009/10) using different feature subsets and compared their RPS (on the remaining seasons). As Table 12.3 shows, each feature set extension leads to a small improvement of the model's performance.

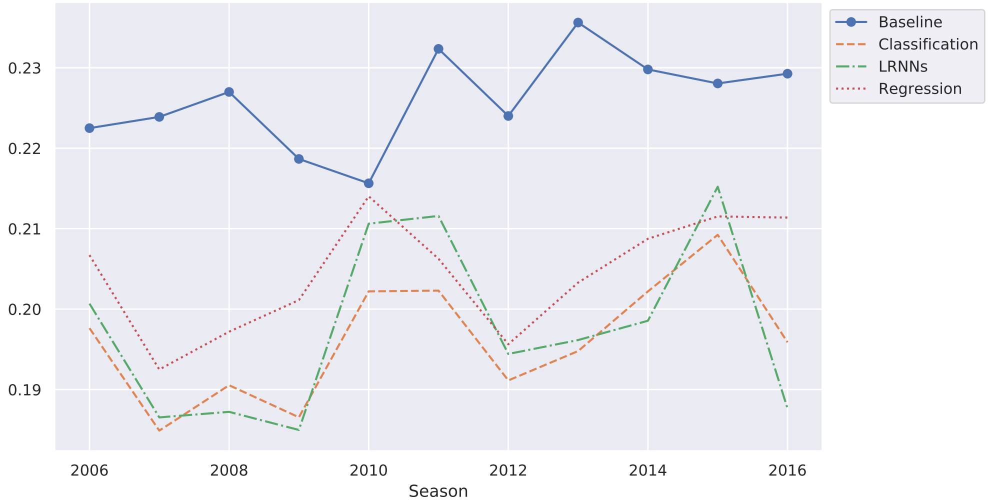

Figure 12.4: Comparison of performance of the learners on English Premier League.

12.6 LRNNS

Training the LRNNs on the whole dataset proved to be too computationally demanding. Therefore, we limited the dataset for evaluating this relational learner to the world's most prestigious English Premier League over the seasons 2006-2016. LRNNs were trained sequentially with a history span of 5 years.

We display the final results in Figure 12.4. All the learners easily pass the natural baseline (mean RPS 0.2260), with LRNNs (0.1976) trailing just closely behind the classification model (0.1961).

12.7 COMPARISON WITH STATE-OF-THE-ART

In this section we compare the selected classification model (Section 10.2) against the state-of-the-art (Section 12.4). To comply with the validation framework used in our experimental review (Section 6) we had to redo the hyperparameter search. The hyperparameter optimization was done by Optuna [2]. The best set of hyperparameters found by Optuna follows: *max_depth* (= 4), *subsample* (= 0.6), *min_child_weight* (= 20) and *colsample_bytree* (= 0.4).

The selected model outperformed the state-of-the-art model by a considerable margin (Table 12.4).

We also examine the similarities between the selected model's predictions and the state-of-the-art models' predictions (Table 12.5). We observe, that the classification model's predictions are quite distinct from both the best statistical model and the best

Table 12.4: Comparison of the selected model against the state-of-the-art.

	xEnt	RPS	Acc.
classification	1.0215	0.2093	48.75
Berrar	1.0246	0.2101	48.54
Bivariate Poisson	1.0251	0.2103	48.58

rating. However, some level of divergency is expected as the model performed much better (Table 12.4).

Table 12.5: Average Jensen-Shannon divergence between the selected model and the state-of-the-art models.

	classification	Berrar	Bivariate Poisson
classification	0.000	0.042	0.043
Berrar	0.042	0.000	0.051
Bivariate Poisson	0.043	0.051	0.000

Finally, we investigate how quickly the model adapts to new data (Figure 12.5). The classification model consistently outperforms the state-of-the-art model. Namely in the beginning of the seasons, the performance gap between the models is noticeably large.

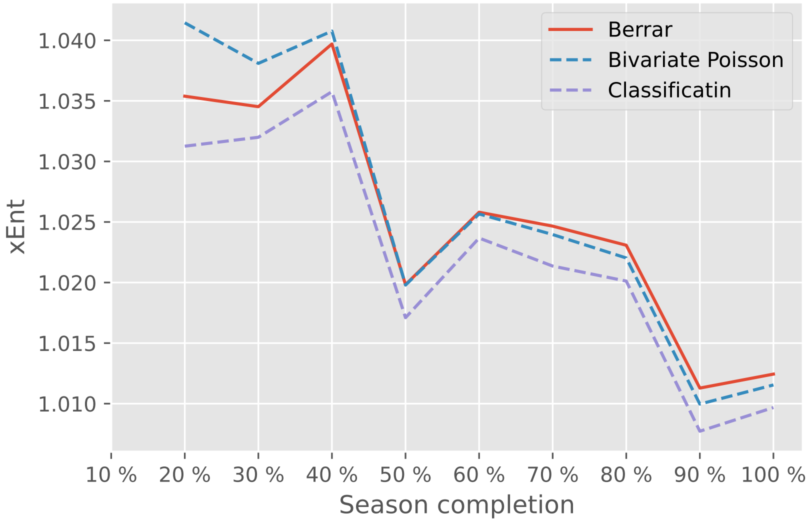

Figure 12.5: Crossentropy of the models as a function of percentage of season completion.

13

DISCUSSION

As expected, all prediction methods fell between the natural lower and upper bounds in terms of the RPS error indicator. Quite surprisingly, the performance of the relational model was not far from that of the best predictor. However, the classifier based on features manually designed with domain insight still turned out to be unmatched in performance. The selected classification model outperformed the state-of-the-art model(s) in all metrics measured. This is a strong result, as while the xgboost model leverages large number of features, all the features were directly derived from the final scores. We illustrate the viability of the feature engineering approach. However, we are aware that adapting the feature engineering process to another domain would be more time-consuming than adapting for example Elo (Section 5.2.1) to that domain. Nevertheless, the results in Table 12.4 indicate much larger improvement than was achieved in recent years in score-based modeling (Section 7). An undisputed advantage of using a machine learning model is that the model can integrate outputs of other models (Section 5).

While we included only the pi-ratings into the features, it is quite likely that adding, for example, the scoring rates from a Poisson model would improve the performance even further. On the other hand, as the statistical models and ratings models were very correlated in their respective groups (Table 7.2), it is possible that inclusion of multiple ratings and scoring rates would not benefit the model. While the improvement over the state-of-the-art is substantial, the bookmakers remain far out of reach (Section 12.3). It seems unlikely that a performance gap of this size could be overcomed by making small incremental changes to the model (i.e. including outputs of other score-based models). This observation calls for the use of more complex models and features or a different approach to beating the market rather than relying on an overall better model.

CRITIQUE OF RPS

The regression model performed rather poorly. This model was intended to accommodate the ordinality of target classes (Eq. 10.2). We analyzed its predictions and indeed, the predicted probabilities were always monotone in the sense that either $p_l \leq p_d \leq p_w$ or $p_l \geq p_d \geq p_w$. However, the best-performing classification model

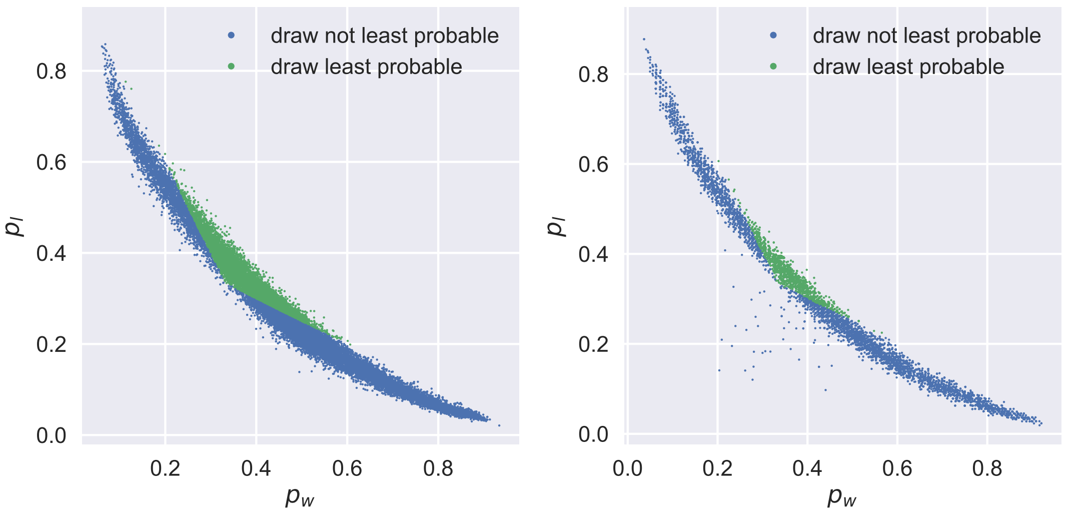

Figure 13.1: Distribution of model's (left) and bookmaker's (right) probability predictions on the p_w-p_l plane. Green dots indicate the cases where p_d is smaller than each of p_w and p_l.

did not adhere to such ordinality. In particular the latter model predicted the draw as the least probable outcome in about 30% of the matches. This calls into question the monotonicity assumption and consequently the suitability of the RPS evaluation measure.

To get further insight into this issue, we analyzed the bookmaker's predictions and found out that the predicted draw probability is smallest among the three predicted probabilities in about 25 % of the matches, further supporting the reasonability of predicting non-monotone probability distributions.

In fact, although the very outcomes of a game are naturally ordered, the corresponding probabilities cannot be reasonably expected to be monotone. Indeed, for a pair of equal-strength teams, the draw is the least probable match outcome. This is given by the prior probability of results which is much lower for draws. As the number of goals in a match increases, the number of possible draws grows linearly, while the number of all possible results grows quadratically. As we can observe from the Figure 13.1, this scenario indeed occurs when the probabilities of the home team winning and losing are close to each other, or in other words, when there is no clear favorite.

Another reason why the draw might be the least probable outcome is that the teams are usually awarded 3 points for a win and 1 point for a draw. In certain situations, a team might consider the draw as a loss of two points instead of a gain of one point. This leads to taking a higher risk during a match when the score is level.

Part IV

BEATING THE BOOKIES

14

INTRODUCTION

In Section 2 we introduced the main parts of a common workflow of a trader t, who relies on a statistical estimator t to predict fair prices R of market opportunities Ω based on some available relevant data $\mathcal{D}^t$, with the resulting estimates of the expected returns $\mathbb{E}_{\mathrm{t}}[\rho]$ being fed into some subsequent portfolio optimization strategy s to produce final wealth allocations $\boldsymbol{f}$. In Part ii we examined several existing estimators for a specific market and in Part iii we introduced our own models. We have seen (Section 12.3) that even the best score-based models trail behind the bookmakers by a large margin. One of the reasons is that the bookmakers do not limit themselves to score-based models and most likely rely on much more detailed data. Such data could provide them with a considerable information advantage (Section 2.4).

In this part of the thesis, we delve deeper into the problem of profiting against the market. To level the playing field w.r.t. the information advantage, we focus our experiments on the domain where comprehensive data are freely available – the basketball, namely the NBA competition.

In Section 15 we summarize the relevant literature. Section 16 provides key insights into the problem of profitability from the market taker's perspective and eposes caveats in common approaches. To tip the scales into our favor, we introduce the notion of *decorrelation* (Section 17). We introduce a novel predictive model in Section 18. In Section 19 we conduct both simulations and experiments on real-world data to put our model and the concept of decorrelation to test.

15

RELATED WORK

Several studies investigated the strategies of bookmakers and bettors. Focusing on the US National Football League (NFL), Levitt [76] traced how odds are set and concluded that bookmakers rely on their ability to outperform an average bettor in outcome forecasting rather than on earning money by balancing weighted wages and profiting from the margin. This hypothesis was subjected to further scrutiny by Paul and Weinbach [99], challenging Levitt's dataset informativeness as consisting only of bets from entry-fee betting tournaments and a limited numbers of participants. However, the conclusions essentially confirmed those of Levitt. Although the hypothesis was not confirmed in basketball [100] using National Basketball League (NBA) data, the disagreement can be explained by the smaller NBA betting market. The recent inquiry [101] into the behavior of bettors with data from NBA and NHL season 2008/09 proposes that most bettors act more like fans than investors. Combined with the conclusion of Levitt [76], this motivates the question whether the bookmaker can be exploited with an emotionless statistical model.

The idea that a statistical model might outperform experts was first tested in Forrest and Simmons [41]. The experts were found unable to process publicly available information efficiently. Signs of using information independent of publicly available data were rare. The study deemed it unlikely that experts would outperform a regression model. Forrest, Goddard, and Simmons [40] challenged the thesis that a statistical model has an edge over tipsters. They examined the performance of a statistical model and bookmakers on 10 000 soccer matches and concluded that bookmakers were on par with a statistical model.

Song, Boulier, and Stekler [117] analyzed prediction accuracy of experts, statistical models and opening betting lines on two NFL seasons. There was a little difference between statistical models and experts performance, but both were outperformed by the betting line. Spann and Skiera [119] compared prediction accuracy of prediction markets, betting odds and tipsters. Prediction markets and betting odds proved to be comparable in terms of prediction accuracy. The forecasts from prediction markets would be able to generate profit against the betting odds if there were not for the high fees. On the other hand, tipsters performed rather poorly in this comparison.

Stekler, Sendor, and Verlander [120] focused on several topics in horse racing and team sports. Forecasts were divided into three groups by their origin – market, models, experts. Closing odds proved to be better predictors of the game outcome than opening odds. The most important conclusion was that there was no evidence that a statistical model or an expert could consistently outperform betting market.

Franck, Verbeek, and Nüesch [42] inspired by results of prediction markets in different domains such as politics, compared performance of betting exchange against the bookmaker on 3 seasons of 5 European soccer leagues. The prediction market was superior to the bookmaker in terms of prediction accuracy. A simple strategy based of betting on the opportunities where the average odds set by the bookmakers were higher than the odds in prediction market was profitable in some cases.

Angelini and De Angelis [5] examined effectiveness of 41 bookmakers on 11 European major leagues over a period of 11 years. Some of the markets turned out to be inefficient, since a trivial strategy of betting on opportunities with odds in certain range led to positive profit.

15.1 PREDICTIVE MODELS

The review Haghighat, Rastegari, and Nourafza [53] of machine learning techniques used in outcome predictions of sports events points out the prevailing poor results of predictions and the small sizes of datasets used. For improving the prediction accuracy the authors suggested to include player-level statistics and more advanced machine learning techniques.

Loeffelholz, Bednar, and Bauer [79] achieved a remarkably high accuracy of over 74% using neural network models, however their dataset consisted of only 620 games. As features, the authors used seasonal averages of 11 basic box score statistics for each team. They also tried to use average statistics of past 5 games and averages from home and away games separately but reported no benefits.

Ivanković et al. [58] used ANNs to predict outcomes of basketball games in the League of Serbia in seasons 2005/06–2009/10. An interesting part of the work was that effects of shots from different court areas were formalized as features. With this approach, the authors achieved the accuracy of 81 %. However, their very specific dataset makes it impossible to compare the results with other research.

Miljković et al. [92] evaluated their model on NBA season 2009/10. Basic box score statistics were used as features, as well as win percentages in league, conference or division and in home/away games. A Naive Bayes classifier in 10-fold cross-validation achieved mean accuracy of 67 %.

Puranmalka [107] used play-by-play data to develop new features. The main reason why features derived from such data are superior to box score statistics is that they include a context. Out of Naive Bayes, Logistic Regression, Bayes Net, SVM and k-NN, the SVM performed best, achieving accuracy over 71 % in course of 10 NBA season from 2003/04 to 2012/13.

Zimmermann, Moorthy, and Shi [134] leveraged multi-layer perceptrons for sports outcome predictions. They proposed the existence of a *glass ceiling* of about 75 % accuracy based on results achieved by statistical models in numerous different sports. This glass ceiling could be caused by using similar features in many papers. They also argued that the choice of features is much more important than the choice of a particular machine learning model.

Vračar, Štrumbelj, and Kononenko [131] made use of play-by-play data to simulate basketball games as Markov processes. Analysis of the results showed that a basketball game is a homogeneous process up to the very beginning and end of each quarter. Modeling these sequences of the game had a large impact on forecast performance. The author saw the application of their model not only in outcome prediction before the game but also in in-play betting on less common bets (number of rebounds/fouls in specific period of the game).

Maymin [86] tested profitability of deep learning models trained on different datasets during the course of a single NBA season. In the paper, positive profits were only achievable with the use of detailed features extracted by experts from video recordings, while models trained using standard box-score statistics terminated with significant loss.

Constantinou, Fenton, and Neil [28] designed an ensemble of Bayesian networks to assess soccer teams' strength. Besides objective information, they accounted for the subjective type of information such as team form, psychological impact, and fatigue. All three components showed a positive contribution to models' forecasting capabilities. Including the fatigue component provided the highest performance boost. Results revealed conflicts between accuracy and profit measures. The final model was able to outperform the bookmakers.

Sinha et al. [115] made use of twitter posts to predict the outcomes of NFL games. Information from twitter posts enhanced forecasting accuracy, moreover, a model based solely on features extracted from tweets outperformed models based on traditional statistics.

15.2 PORTFOLIO OPTIMIZATION

The approach of splitting the trader's workflow into the two steps of predictive modeling and investment optimization has a long tradition, and has been exploited in absolute majority of works [39, 59, 90, 95, 102, 124], with some notable exceptions [50, 75]. Extracting the parameter estimation out of the portfolio optimization problem then enabled the respective economic research to thrive in an isolated mathematical environment, giving rise to the frameworks of Markowitz [85] (Section 2.7.2) and Kelly [63] (Section 2.7.3), and their many successors [19, 71, 95, 124, 132]. While widely adopted, the optimality of the resulting portfolios is based on rather unrealistic assumptions, which has been progressively criticized by many [55, 81, 90, 103, 111, 112]. From the perspective studied in this thesis, the main underlying issue is the separation from the problem of estimation of the return (price) parameters, which are simply assumed at input. The resulting issues with uncertainty in the portfolios are then typically mitigated with additional practical methods [60, 80, 95]. There are also some principled approaches to tackle the input parameter uncertainty, such as considering the portfolio optimization problem within the framework of Bayesian decision making [10, 20, 24] or distributionally robust setting [14, 123]. However, to our best knowledge, all of these methods are aimed at mitigating the additional (structural) risk, stemming from the uncertainty in the input parameters, rather than increasing the profits.

16

PROBLEM INSIGHTS

In this Section, we provide some key insights into the problem of profitability from the perspective of the predictive model t.

Let us briefly recall the problem setup. We generally consider the problem of profiting from the trader's t perspective as a stochastic game against the market maker m. The market maker m uses an estimator m to continuously price the incoming opportunities Ω. Following some investment strategy, the trader t then takes particular α and β actions (allocations) upon these opportunities Ω, based on his/her own estimates produced by t. Given some distribution of market opportunities P_Ω, we can then set up the game in terms of three random variables R, M, T corresponding to the fair price, market maker's, and market taker's estimates, respectively. The goal of the trader (as well as the market maker) is then to maximize his/her expected (long-term) profits W as measured by some utility u underlying the chosen strategy s.

16.1 FROM ACCURACY TO PROFIT

The key issue with the optimal investment strategies based on portfolio optimization (Section 2.7) is that they are inherently relying on accurate estimates of the asset returns. Their performance is then directly stemming from the quality of these estimates – the better the estimates, the higher the utility of the portfolio can be achieved in general. While this holds to an extent for most of the common portfolio optimization strategies, it is best demonstrated on the optimal investment approach of Kelly.[1]

For simplicity of demonstration, let us consider an idealized case of a betting market with no spread (Section 2.3) on the market maker's odds. Recall that the Kelly strategy is to find wealth fractions $\boldsymbol{f}$ so as to

$$\begin{aligned} \underset{\boldsymbol{f}}{\text{maximize}} \quad & \mathbb{E}_R\left[\log\left(\boldsymbol{f}^T \cdot (1+\boldsymbol{\rho})\right)\right] = \sum_{i=1}^{n} r_i \log\left(f_i \cdot \frac{1}{m_i}\right) \\ \text{subject to} \quad & \sum_{i=1}^{n} f_i = 1,\ f_i \geq 0 \end{aligned} \tag{16.1}$$

[1] The correspondence between Kelly and MPT is shown in Section 2.7.3.

Note that in this idealized case, we calculate the expectation of returns w.r.t. the true distribution $P_\Omega(R)$. It can then proved [29] that the solution to this constrained optimization problem yields

$$\boldsymbol{f}^* = \boldsymbol{r} \tag{16.2}$$

i.e. the optimal fraction of wealth f_i to invest in each outcome (opportunity) ω_i is directly equal to the underlying fair price (probability) r_i. Interestingly, we can see that in this case, the optimal strategy for the investor is to completely ignore the market pricing m_i and focus solely on having the fair prices values predicted correctly, in which case he/she is guaranteed the maximal possible long term profits. This is commonly known amongst Kelly practitioners as "betting your beliefs". Note that this strong result was derived from $\mathbb{E}_R$ and thus it only holds if the true distribution $P_\Omega(R)$ is known or, more precisely, if the error in its estimate via $T = \hat{R}$ as measured through the Kullback-Leibler divergence (Section 2.5) is zero [29]:

$$D_{KL}(R||T) = -\sum_{i=1}^{n} r_i \cdot \log \frac{t_i}{r_i} = 0 \tag{16.3}$$

Naturally, it is close to impossible to estimate the true distribution $P_\Omega(R)$ perfectly in practice. Let us thus extrapolate into more practical settings by relaxing the condition into a non-zero $D_{KL}(R||T)$. Given that the optimal fractions $\boldsymbol{f}$ should be equal to the true outcome probabilities $\boldsymbol{r}$, let us substitute back into the long term growth rate of wealth which Kelly seeks to maximize as

$$W_\mathsf{G} = \sum_{i=1}^{n} r_i \log \left(t_i \cdot \frac{1}{m_i} \right) \tag{16.4}$$

Now, following the proof from [29], this can be rewritten into

$$W_\mathsf{G} = \sum_{i=1}^{n} r_i \log \frac{t_i}{r_i} + \sum_{i=1}^{n} r_i \log \frac{r_i}{m_i} \tag{16.5}$$

and consequently, using the formula for KL-divergence (Equation 2.8), back into

$$W_\mathsf{G} = D_{KL}(R||M) - D_{KL}(R||T) \tag{16.6}$$

showing the important insight that, for Kelly, the growth of wealth of the trader is directly equal to the difference in quality of his/her estimates T over the market prices M in terms of KL-divergence from the fair prices R. Consequently, positive returns can only be achieved iff the model of the trader achieves a lower cross-entropy error than the market $XENT_\Omega(R, M) < XENT_\Omega(R, B)$ (Section 2.5). Given the information theoretic interpretation of the relative entropy [69], this is sometimes referred to as the aforementioned "information advantage" of the trader over the market maker.

IMPLICATIONS Note that this result was derived with the assumption of seeking growth-optimal investments, and its extrapolation beyond that setting may lead to wrong conclusions. Particularly, it is true that one does need a better[2] model to make positive profits if committed to invest optimally with Kelly. However, this does not imply that one needs a better model to make positive profits if one does not require the growth optimality.

The constraint for better model accuracy in classic portfolio optimization techniques is then inherently connected to the notion of risk (Section 2.6), which is embedded together with expected returns into the same quantity being optimized. For instance, in the Markowitz's model, it is easy to show that even negative return portfolio may be preferred to positive returns should the latter be associated with higher variance. From the Kelly's perspective, overvalued positive return estimates may actually lead to negative growth due to overbetting (Table 16.1), and it is also commonly necessary to allocate certain amount of wealth onto opportunities (outcomes) with negative returns to achieve optimal growth [126].

Consequently, wrong assessment of the fair prices associated with either of such opportunities can lead to inappropriate (over-)investments, resulting into a negative overall profit, even in situations where positive returns could be generally achieved otherwise.

16.2 THE ESSENCE OF PROFIT

While the performance of common portfolio optimization strategies is tightly bound to the accuracy of price predictions (Section 16.1), we argue that accuracy is not essential for profitability in general. This is best demonstrated by taking the, sophisticated but questionable, notions of risk out of the optimization scope, resorting back to simple strategies such as the uniform investments (Section 2.7.1). Consequently, one can simply base profitability directly on the ability to correctly detect opportunities with positive expected returns (Section 2.6). Note now that whether the expectation from an opportunity ω_i is deemed positive depends purely on the comparison between t_i and m_i. This boils down to the renown "buy low, sell high" policy to trade mispriced selections simply as:

$$m_i \neq t_i \begin{cases} m_i < t_i \implies \alpha = \text{back (buy) } \textit{assumed} \text{ underpriced selection} \\ m_i > t_i \implies \beta = \text{lay (sell) } \textit{assumed} \text{ overpriced selection} \end{cases} \tag{16.7}$$

Naturally, to asses the actual return from a trade, the true selection value r_i needs to be accounted for. The actual return from a supposedly profitable opportunity can then be defined as

[2] We note we do not distinguish between $XENT$ and other measures of model accuracy here for simplicity.

$$\mathbb{E}_R[\rho_i] = \begin{cases} \frac{r_i}{m_i} - 1, & \text{if } m_i < t_i \quad \text{(backing selection)} \\ \frac{1-r_i}{1-m_i} - 1, & \text{if } m_i > t_i \quad \text{(laying selection)} \end{cases} \tag{16.8}$$

Note that the true expected return $\mathbb{E}_R[\rho_i]$ can clearly be negative and that its absolute value is not dependent of the model's estimate t_i. However, the ability to correctly recognize the profitable opportunities through the $m_i \lessgtr t_i$ comparison is naturally dependent on the ordering of the t_i estimates w.r.t. m_i and r_i. Note nevertheless that this comparison-based quality is very different from the accuracy-based reasoning. Consequently, even if $err(\mathrm{t}) > err(\mathrm{m})$, a consistent profit can still be made, as we demonstrate through the following simple examples.

Example 16.2.1. Assume fair price of a selection to be 0.6, with the bookmaker m estimating it at $m_i = 0.5$, with the corresponding odds set up to 2.0, and the bettor t estimate being at $t_i = 0.9$. Clearly, the bettor's estimate is more erroneous here (e.g. $XENT(t_i) > XENT(m_i)$). Nevertheless he/she has no choice but to use his/her estimate to asses the return on investment, which he/she *correctly* estimates as being *positive* ($\mathbb{E}_\mathrm{t}[\rho] = \frac{0.9}{0.5} - 1 > 0$). Despite being very wrong numerically with his/her expectation of a 80% ROI, by betting a unit of wealth, he/she can still expect to obtain the actual positive ROI of 20%.

Note that the trader's t estimates t_i in these examples could have been set arbitrarily larger (within the respective domain), making the corresponding model t arbitrarily bad by the standard error measures.

Definition 16.2.1. Followingly, let us define a more relaxed, necessary condition of *essential profitability* of a model t simply as the consequent existence of one of the following market opportunities ω_i in $P_\Omega(R, M, T)$:

1. the market undervalues the fair price, and the model estimates a higher value than the market,
   i.e. $m_i < r_i \wedge t_i > m_i$

2. the market overvalues the fair price, and the model estimates a lower value than the market,
   i.e. $m_i > r_i \wedge t_i < m_i$

Using sufficiently conservative (small) wealth allocations, investments into either of these cases will lead to systematic profits of the market taker in the long run. On the contrary, no investment strategy can lead to positive profits without such opportunities in the portfolio.

Nevertheless the overall profitability of a model t will naturally depend on the relative occurrence of such ω_i's in the actual market distribution P_Ω (Definition 2.5.1). Let us now generalize the essence of profitability, from reasoning about the necessary relationships between individual r_i, m_i, t_i estimates, to the properties of the whole market distribution $\mathrm{P}_\Omega(R, M, T)$. Following on the aforementioned "buy low, sell high" strategy with uniform investments, the expected profitability from a market distribution P_Ω is clearly

$$\mathbb{E}_{P_\Omega}[\rho] = \sum_i \mathbb{E}_R[\rho_i] \cdot \mathrm{P}_\Omega(r_i, m_i, t_i) \tag{16.9}$$

We already know that the fair price r_i is a principally unknown random variable and one can thus never perfectly assess the true return ρ_i from any trade in advance, for which we resort to an estimate $\hat{\rho}_i$. However, the market distribution $\mathrm{P}_\Omega(R, M, T)$ here is also principally unknown, for which one again needs to rely on statistics while estimating it from historical data as $\hat{\mathrm{P}}$. Consequently, one can estimate the essential profitability of a model t w.r.t. market pricing m as

$$\mathbb{E}_{\hat{P}}[\widehat{\rho}] = \sum_i \mathbb{E}_T[\rho_i] \cdot \hat{\mathrm{P}}(t_i, m_i, t_i) \tag{16.10}$$

As with any investment strategy, the calculated expected returns can be very different from the actual return distribution $\mathbb{E}_{\hat{P}}[\widehat{\rho}] \neq \mathbb{E}_{P_\Omega}[\rho]$, depending on the properties of the estimates (Section 2.6). However, profitability of the simple unit stake strategy leads to a much more relaxed and robust condition on the model quality, which is what we exploit to yield positive profits even with estimators of inferior predictive performance.

DRAWBACKS We note that the standard approach of focusing on predictive accuracy has many advantages, such as being naturally less noisy and more interpretable, and should be preferred when profitability is not the target [133]. We also acknowledge that by deflecting from the accuracy-based view and focusing merely on the essential profitability with the unit investments, we downplay the role of explicit optimization of growth and risk in the formal strategies of Kelly and Markowitz, respectively. Nevertheless, as discussed in the respective Sections 2.7.3 and 2.7.2, these formal notions are based on rather unrealistic (wrong) assumptions, which is why additional risk management practices, such as the fractioning (Section 2.7.3), need to be commonly employed with the strategies anyway [82, 83, 127]. Consequently, sacrificing formal optimality w.r.t. unrealistic objectives in order to transition from negative to positive profits does no seem that big of a sacrifice.

Table 16.1: All possible orderings of the fair price (r_i), market maker's (m_i), and trader's estimates (t_i), with the implied trading decisions and resulting profitability. Additionally, the relative size of the implied Kelly fraction is indicated.

values ordering	decision	profit	Kelly
$r_i < t_i < m_i$	lay	$m_i > r_i$	overbet
$r_i < m_i < t_i$	back	$r_i < m_i$	overbet
$t_i < r_i < m_i$	lay	$m_i > r_i$	underbet
$t_i < m_i < r_i$	lay	$m_i < r_i$	underbet
$m_i < t_i < r_i$	back	$r_i > m_i$	underbet
$m_i < r_i < t_i$	back	$r_i > m_i$	overbet

16.3 MARKET TAKER'S ADVANTAGE

The market maker's advantage (Section 2.3) is a well-worn concept. However, there is also an advantage of the market *taker* which is rarely discussed explicitly, but is essential to the traders profitability. While the market maker has the obligation to continuously *quote* price of both sides of the market (Section 2.5), the taker has the crucial liberty to *select* only those of the resulting opportunities deemed profitable. That is he/she is to decide whether and which side of the market to trade once the market maker's prices have been laid out. As the second player, the difficulty of his/her task is reduced from the correct price estimation to the estimation of the market price error direction. While this might seem as a similarly difficult problem, the latter is a considerably easier task.

For demonstration, consider the three values r_i, m_i, t_i of the fair price, market maker's, and trader's estimates, respectively, to be laid out completely at random, yielding a uniform market distribution $P_{\mathcal{U}}$ where $R, M, T \sim \mathcal{U}^3$. The possible situations that emerge from the ordering of r_i, m_i, t_i in such a setting are displayed in Table 16.1. Note that the 6 particular orderings are distributed evenly in a uniform distribution. Since neither of the estimators possesses any information w.r.t. R, both M and T clearly perform equally by the means of arbitrary statistical estimation measures (Section 2.5). While one might thus expect this to be a neutral trading setting for both the sides, interestingly, the market taker t would already be able to make a substantial profit with uniform investments by correctly identifying 2/3 of the profitable opportunities.

Intuitively, this demonstrates a simple fact that it is generally more likely to overestimate an undervalued estimate than to further underestimate it, i.e.

$$m_i < r_i \implies P_{\mathcal{U}}(m_i < t_i) > P_{\mathcal{U}}(t_i < m_i) \tag{16.11}$$

and vice versa for an overvalued estimate:

$$m_i > r_i \implies P_{\mathcal{U}}(m_i < t_i) < P_{\mathcal{U}}(t_i < m_i) \tag{16.12}$$

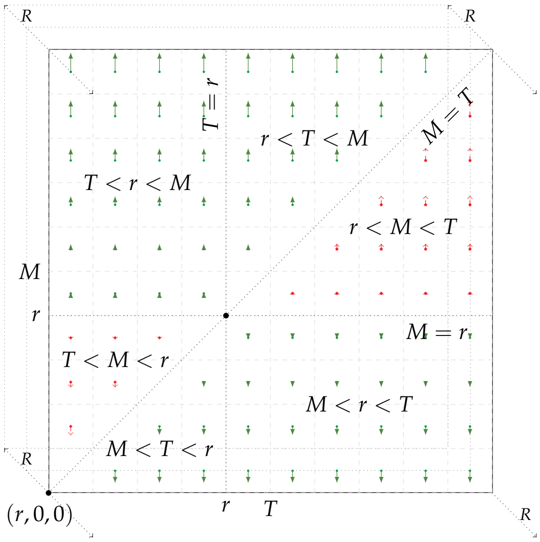

Figure 16.1: A 2D projection of the $\mathrm{P}(R, M, T)$ distribution onto $\mathrm{P}(M, T|R = r)$ with visualization of essential profitability (Section 16.2) of the individual point-estimates as a vector field. Green color (solid arrows) denotes positive returns and red (dashed arrows) negative returns, respectively, while the length of each vector corresponds to the ROI in an idealized two-sided market (Section 2.6).

Note, importantly, how this property of the completely uninformative $\mathrm{P}_{\mathcal{U}}$ is conveniently aligned with the essential profitability of the trader's model (Definition 16.2.1). The concrete proportions of the individual situations will naturally depend on the particular distribution P_{Ω}, nevertheless the property holds very generally for unskewed distributions with unbiased estimators (Section 16.4). Note the difference from the standard model accuracy measures which would all evaluate both models equally in $\mathrm{P}_{\mathcal{U}}$. Nevertheless from the perspective of profitability, the situation is very different since, as opposed to the market maker, the market taker is not penalized for estimation errors in these two situations that emerge more often than not. Consequently in $\mathrm{P}_{\Omega} = \mathrm{P}_{\mathcal{U}}$, the trader is in an inherent advantage of $2 : 1$, which can be directly turned into the corresponding profits.

It is perhaps more instructive to demonstrate the concept on a particular level of fair prices. Without loss of generality, let us consider all selections with a fair price r_i being traded in an ideal two-sided market (without spread). We can then plot a 2D projection of the $\mathrm{P}(R, M, T)$ market distribution by conditioning it as $\mathrm{P}(M, T|R = r)$, and visualize the essential profitability (Definition 16.2.1) of the corresponding subregions of the distribution. The result is displayed in Figure 16.1. We can observe that the distribution of the profitable regions (green, solid arrows) is clearly in favor of the trader, and that the potential returns progressively increase with the error of the market maker, i.e. the distance of $M = m$ from r.

Note that here we did not yet assume any particular (non-uniform) distribution of estimates, and this inherent advantage of the trader is thus completely oblivious of any information advantage as well as any other property of the price estimators. Rather, it stems purely from the unequal roles of the market maker m and taker t. Consequently, should they, e.g., switch roles with the same estimation models ($m \leftarrow \mathtt{t}, t \leftarrow \mathtt{m}$), the advantage would stay exactly the same on the side of the market taker t.

16.4 DISTRIBUTION OF ESTIMATES

While demonstrating the concept of the trader's advantage, the uniform $\mathrm{P}_{\mathcal{U}}(R, M, T)$ distribution of estimates with completely uninformed players from Section 16.3 seems rather unlikely in practice. In real world markets, the values R, M, T are not going to be independent, but rather correlated with each other, since M and T are typically based on similar information sources and both try to model R with similar techniques (Section 2.4). Let us now review common properties of the more realistic market distributions of these estimates.

BIAS The market makers are typically very good at being close to the fair price, and we can assume their estimates M to be *unbiased* w.r.t. R, or, more formally:

$$\mathbb{E}_{\mathrm{P}}(B) - R = 0 \tag{16.13}$$

which means that they are not systematically deviating when measured against the fair prices alone. Should a market maker be biased in this manner, it would be extremely easy to exploit him/her merely by trading all opportunities on the corresponding side of the market. Moreover we can reasonably assume him/her to be point-wise unbiased at each particular price level r, i.e.

$$\mathbb{E}_{\mathrm{P}}(B|R = r) = r \tag{16.14}$$

If that was not the case, the market maker would again be easily exploitable by correspondingly trading all possible selections within a certain price range $r \pm \delta$, i.e. by backing all selections in price ranges where the maker m systematically undervalues the selections, and vice versa for laying in overvalued price ranges. One can typically check from historical data that the market makers are not biased in this simple manner in any reasonably efficient market. Note that the unbiasedness is only one of the conditions for a fully efficient market (Section 2.2). There is generally no reason for the market taker t to be biased in this trivial way either, unless a systematic error is present in his/her model $\mathtt{t}$, or he/she reversely reflects the market maker's bias to exploit it.

VARIANCE Given the assumption that the models behind M and T are both unbiased estimators of R, we can now focus merely on their (co-)variances. It is a common practice in statistical estimation of functions for one to look for an estimator with the smallest variance among the class of unbiased estimators. Since we are working with function estimators, we are not interested in the total variance of the estimations T, which includes variance due to variations in R itself as

$$\mathrm{Var}[T] = \mathbb{E}_R[\mathrm{Var}[T|R]] + \mathrm{Var}_R \mathbb{E}[T|R] \tag{16.15}$$

but merely in the first term capturing the expected variance left w.r.t. predicting R. Given the assumption of the point-wise unbiased estimates, the conditional variances w.r.t. R are then equal to the covariances of the models, i.e.

$$\mathrm{Cov}[M, R] = \mathrm{Var}[M|R] \quad \text{and similarly} \tag{16.16}$$

$$\mathrm{Cov}[T, R] = \mathrm{Var}[T|R] \tag{16.17}$$

Given the unbiasedness, these co-variances then directly reflect the quality (accuracy) of the underlying models of the market maker ($\mathrm{Cov}[M, R]$) and market taker ($\mathrm{Cov}[T, R]$), respectively.

The last degree of freedom in terms of covariances in P is the relationship between M and T, i.e. $\mathrm{Cov}[M, T]$ ($\mathrm{Cov}[M, T|R]$). While the other two covariances have the clear introduced interpretation, the $\mathrm{Cov}[M, T]$ is more intriguing, but is also essential to the proposed profitability of inferior predictive models (Section 16.2.1). As we have seen in the case of the uniform market distribution $\mathrm{P}_{\mathcal{U}}$, where both the players possess the same amount of information w.r.t. the fair price, corresponding to the same accuracies of $\mathtt{m}$ and $\mathtt{t}$, the trader t is always in advantage. However not all distributions with equally informed players are as such, as demonstrated by the following example.

Example 16.4.1. Assume a scenario where both the trader t and market maker m possess the exact same model. Clearly, their information value, accuracy and all statistical measures will be exactly the same, just as in the case of the uniform distribution $\mathrm{P}_{\mathcal{U}}$. Nevertheless, the trader will not be in an advantageous position anymore. Since all his/her estimates coincide with the market price $\forall i : t_i = m_i$, it is not possible to detect any profitable opportunities where $r \notin (m_i - \epsilon, m_i + \epsilon)$, even if they exist in the market distribution P_{Ω}. Hence, the profitability of the trader in this case is clearly zero.

From the statistical viewpoint, one can note an underlying difference between the two example distributions in the third covariance term $\mathrm{Cov}[M, T]$. Whereas in the uniform distribution, the two variables were completely independent, i.e. $\mathrm{Cov}[M, T] = 0$, here they are equal and thus display maximal possible covariance ($\mathrm{Cov}[M, T] = \sigma^2$). While this anecdotal reference to the connection between profitability and $\mathrm{Cov}[M, T]$ is rather informal, we analyse it in detail in the next Section 17.

16.5 CONFIDENCE THRESHOLDING

We also explored a modification applicable to each of the betting strategies, in which only high-confidence predictions are considered. More precisely, a probability estimate $\hat{p}_i$ is passed to the betting strategy if and only if

$$|\hat{p}_i - 0.5| > \phi.$$

The reasoning behind this thresholding is that we want to remove the games where the model is very indifferent about the favorite. Although being in principle valid for the strategy, our assumption is that probabilistic predictions around 0.5 are typically more imprecise than predictions of higher confidence. This is especially true for the proposed models trained with gradient descent techniques over logistic sigmoid output which is indeed most sensitive at that point.

17

INCREASING PROFIT THROUGH DECORRELATION

Recall that we have a two-sided market with opportunities Ω of some fair price (resulting in R), being priced by the market maker m as M and taker t as T, resulting into some market distribution of estimates $\mathrm{P}_\Omega(R, M, T)$ (Definition 2.5.1). Let us now consider the context of the common properties (Section 16.4) of such market distributions P_Ω, allowing assessments of model performance in terms of expected returns $\mathbb{E}_{\mathrm{P}_\Omega}(\rho)$ w.r.t. the distribution $\mathrm{P}_\Omega(R, M, T)$. Particularly, we will explore the aforementioned statistical relationship between the market maker m and taker t. To further standardize the relationship study, i.e. to take the individual variances of M and T out of scope, we now switch from covariance $\mathrm{Cov}[T, M|R]$ to correlation $\mathrm{Corr}[T, M|R]$.

Definition 17.0.1. We use the term *decorrelation* to refer to the concept of decreasing the partial correlation $\mathrm{Corr}[T, M|R]$ between a price estimator t of the trader and the market maker m w.r.t. the fair pricing function r across opportunities $\omega_i \in \Omega$ endowed with some market distribution P_Ω (Definition 2.5.1).

The main goal of this section is to show that enforcing smaller[1] partial correlation $\mathrm{Corr}[T, M|R]$ of a price estimator t generally increases its essential profitability (Definition 16.2.1) within common market distributions, particularly for estimators t that are inferior to the market maker m.

17.1 UNBIASED ESTIMATORS

Let us first consider the common setting of unbiased price estimators, the reasoning behind which was introduced in Section 16.4.

Theorem 17.1.1. The essential profitability (Definition 16.2.1) of an unbiased estimator t with the lowest partial correlation $\mathrm{Corr}[T, M|R = r] = -1$ with the market m is maximal. Consequently, no deviation from such a model t can thus increase the profitability further.

Proof. Now for $\mathrm{Corr}[T, M|R] = -1$ and an arbitrary r, the probability distribution $\mathrm{P}(T, M|R = r)$ collapses into a linear function $(t; a, b) \mapsto m$ of the form

[1] Note that utilize the term "decorrelation" to refer to decreasing the correlation even below zero.

$$m = a \cdot t + b \text{ where } a = \frac{\mathrm{Cov}[M, T|r]}{\mathrm{Var}[T|r]} \text{ and } b = (1 - a) \cdot r \tag{17.1}$$

where the b is set so that the mean values $\mathbb{E}_{\mathrm{P}(T,M|R=r)}[M] = \mathbb{E}_{\mathrm{P}(T,M|R=r)}[T]$ of both the marginals $\mathrm{P}(M|r)$ and $\mathrm{P}(T|r)$ are at r (unbiased). The mean of the distribution $m = t = r$ thus lies on the line:

$$r = a \cdot r + (1 - a) \cdot r \tag{17.2}$$

Clearly, since $\mathrm{Corr}[T, M|r] < 0$, we have also $\mathrm{Cov}[M, T|r] < 0$ implying a negative slope $a < 0$ of the function line. There are consequently only 2 possible price estimate orderings (regions in Figure 16.1) for all Ω, both of which are profitable, as follows:

1. $T < r < M$ implying returns $\frac{1-r}{1-m} - 1 > 0$
2. $M < r < T$ implying returns $\frac{r}{m} - 1 > 0$

Note again that the individual values of returns do not depend on the value of T but merely on the value order. From the assumed role of the trader t, the only possible changes to the distribution (and profit) can be made via changes in his/her model t estimates T. However, any potential deviation from this distribution (line) with $\mathrm{Corr}[M, T|R] = -1$ can only result into one of the following ordering (region) transitions:

1. $T < r < M$ can change to either:
   a) $r < T < M$ implying *no change* in the returns $\frac{1-r}{1-m} - 1 > 0$
   b) $r < M < T$ implying *decrease* in the returns to $\frac{r}{m} - 1 < 0$
2. $M < r < T$ can change to either:
   a) $M < T < r$ implying *no change* in the returns $\frac{r}{m} - 1 > 0$
   b) $T < M < r$ implying *decrease* in the returns to $\frac{1-r}{1-m} - 1 < 0$

ergo no deviation from $\mathrm{Corr}[M, T|R] = -1$ can increase the profitability any further. □

Note that this also means that we cannot increase the returns even via transition into a perfect estimator with both $\mathrm{Bias}_R[T] = \mathrm{Var}[T|R] = 0$, i.e. a perfect model $\mathrm{P}(T = r|R = r) = 1$ which always returns the correct answer in a deterministic fashion (and for which the partial correlation would be undefined). This means that for the purpose of the essential profit generation (Definition 16.2.1), the variance of the model no longer acts as an error. Note this is in direct contrast to a correlated investor who, given a variance higher than the market maker, is doomed to obtain completely

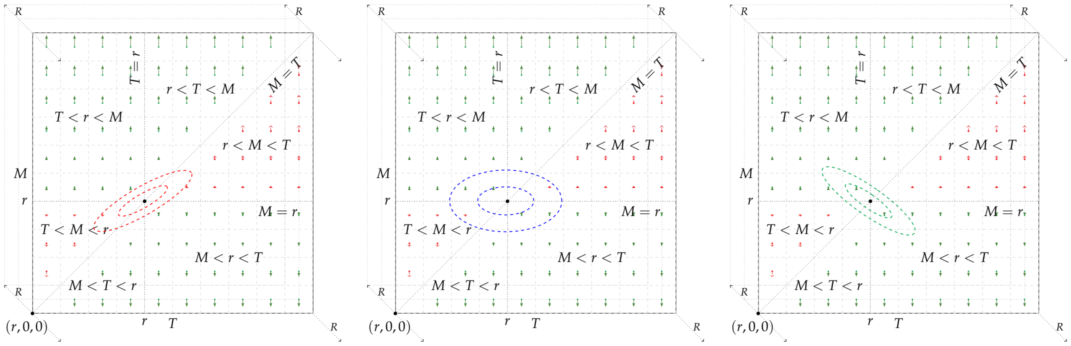

Figure 17.1: A sketch of the profitability regions from Figure 16.1 with an example distribution of estimates density for the case of unbiased estimations of M and T w.r.t $R = r$, where $\text{Var}[T|R] > \text{Var}[M|R]$, corresponding to a common situation in practice (Section 16.4, 19). The effect of $\text{Corr}[T, M|R]$ on profitability is demonstrated on a highly correlated model (left, red), an independent model (middle, blue), and a highly negatively correlated model (right, green). By decreasing the correlation, progressively larger parts of the profitable regions of the distributions (green) are being covered, reflecting the corresponding increase in returns.

negative returns. [2] A visualization of the correlation effect on the returns, with an example elliptical market distribution for the given setting, is depicted in Figure 17.1.

17.2 BIASED ESTIMATORS

While common, the assumption of point-wise unbiased estimators might be seen as too strict in practice. The market maker is very unlikely to be overly biased, due to its constant exposure to the traders, who would likely exploit such easy opportunities (Section 16.4). Nevertheless the market takers are generally free to come up with all sorts of models. Let us briefly review the situation where the trader's model t, which we seek to optimize, is more biased than m w.r.t. R, i.e. $\text{Bias}_R[T] > \text{Bias}_R[M]$.

[2] We also note that the transition between the two corner cases is somewhat smooth w.r.t. the returns for common market distributions, such as the elliptical distributions used for visualization. However, we acknowledge that we do not present a formal proof of monotonicity of this property.

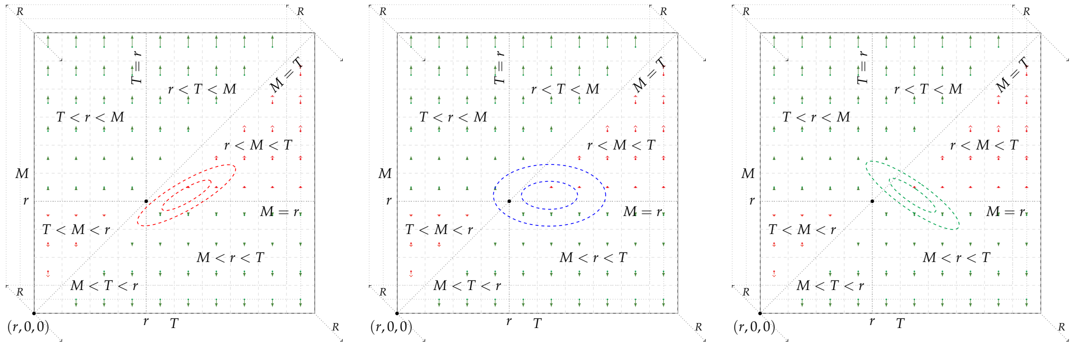

Figure 17.2: A sketch of the distribution of estimates density from Figure 16.1 for the case where the trader T is biased more than the market maker M w.r.t $R = r$. The decorrelation technique still commonly helps as larger parts of the profitable regions of the distribution are being covered by decreasing $\mathrm{Corr}[T, M|R]$.

In this setting, one can craft counterexamples showing that T with $\mathrm{Corr}[T, M|R] = -1$ is not universally the most profitable model t anymore.[3] While not the universally best possible model, a decorrelated T will still perform very well in practice here. Particularly, it will be consistently better than a highly correlated model, even if the latter has a lower variance $\mathrm{Var}[T|R]$. Interestingly, it is also better than a model with zero variance (vertical line), i.e. given some bias, we are able to turn the variance into an advantage by decreasing correlation with M. Lastly, the minimal correlation will be also typically better than no correlation for common elliptical or uniform conditional distributions. A visualization of this setting, where $\mathrm{Bias}_R[T] > \mathrm{Bias}_R[M]$, is displayed in Figure 17.2 for an example elliptical distribution. Note that the same reasoning is also applicable to cases where the $\mathrm{Bias}_R[T] = \mathrm{Bias}_R[M]$.

17.3 HAVING A SUPERIOR MODEL

The primary motivation behind the concept of lowering the correlation with the market is to make profits with models of *inferior* quality, which would commonly yield negative profits otherwise. We argued that such a situation is common, since the market (maker)

[3] For instance, consider a distribution where the model t is biased w.r.t. R by some δ as

$$T = \begin{cases} M - 1.1 \cdot \delta, & \text{if } \; T > r \\ -M + 1.1 \cdot \delta, & \text{if } \; T < r \end{cases} \tag{17.3}$$

This δ-biased model t is set to make maximal possible profit, and decreasing its correlation with m will actually hurt its performance. However, achieving such a distribution of estimates $\mathrm{P}_\delta(R, M, T)$ is close to impossible in practice, as it is carefully crafted w.r.t. the unknown value of r. Consequently, this scenario is highly unstable w.r.t. r as well as variance of T, a decreasing of which will paradoxically lead to complete loss of all returns.

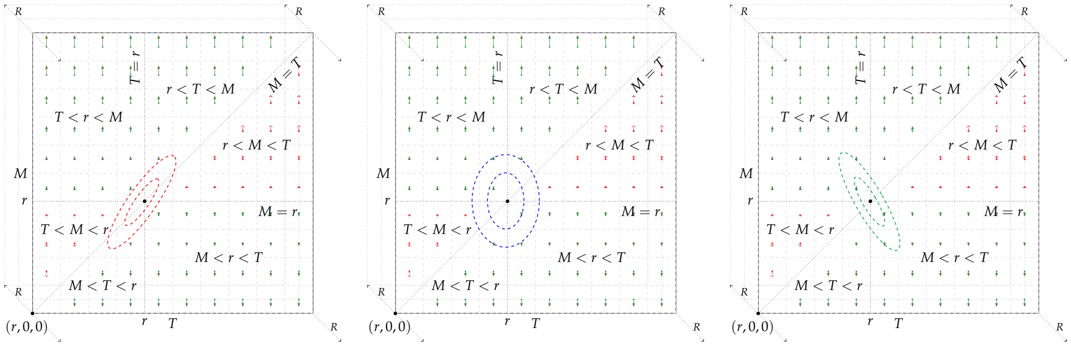

Figure 17.3: A sketch of the distribution of estimates density from Figure 16.1 for the case of unbiased estimators M and T w.r.t $R = r$, where $\mathrm{Var}[T|R] < \mathrm{Var}[M|R]$, corresponding to a superior estimation model t of the trader. For a superior model in terms of the conditional variance, the correlation with the market does not pose such a problem as the model is already profitable, and decreasing it might actually hurt the performance, nevertheless, a model with the lowest covariance still generally provides the highest profitability.

price tends to be a very good estimate of the fair price in any fairly efficient market (Section 16.4). Nevertheless, for completeness, let us consider the opposite case of having a superior model, i.e. a situation where $err(\mathsf{t}) < err(\mathsf{m})$. Following the statistical decomposition of estimation errors in terms of bias and variance (Section 2.5), let us separately consider two cases of such superiority through a model with (i) lower variance and (ii) lower bias.

SUPERIOR VARIANCE For a superior model T with a lower conditional variance $\mathrm{Var}[T|R] < \mathrm{Var}[M|R]$, the concept of decorrelation (Definition 17.0.1) no longer works as a consistent profit enhancement, even for common, realistic market distributions. Particularly, decreasing the correlation $\mathrm{Corr}[T, M|R]$ can actually decrease the returns in many scenarios. We again depict the concept on an example elliptical distribution in Figure 17.3. The variances of the estimators are exactly opposite to those from Figure 17.1. We can see that the returns with an independent model $\mathrm{Corr}[T, M] = 0$ are lower than those of a highly correlated model $\mathrm{Corr}[T, M|R] = 1$. Note however that the decrease of returns in this case is smaller than the increase in the opposite case (i.e. shift from $\mathrm{Corr}[T, M|R] = 1$ to $\mathrm{Corr}[T, T|R] = 0$ in Figure 17.1). Finally, the profits of $\mathrm{Corr}[T, M|R] = -1$ are still maximal.

SUPERIOR BIAS We have argued for the practical necessity of a market maker not to be systematically biased in Section 16.4, which leaves a little space for the trader to beat the market in terms of $\mathrm{Bias}_R[T] < \mathrm{Bias}_R[M]$. Nevertheless it should be acknowledged that in the unlikely case that the market maker indeed is more biased than the trader, the concept of decorrelating the estimates for increased profits breaks down severely.

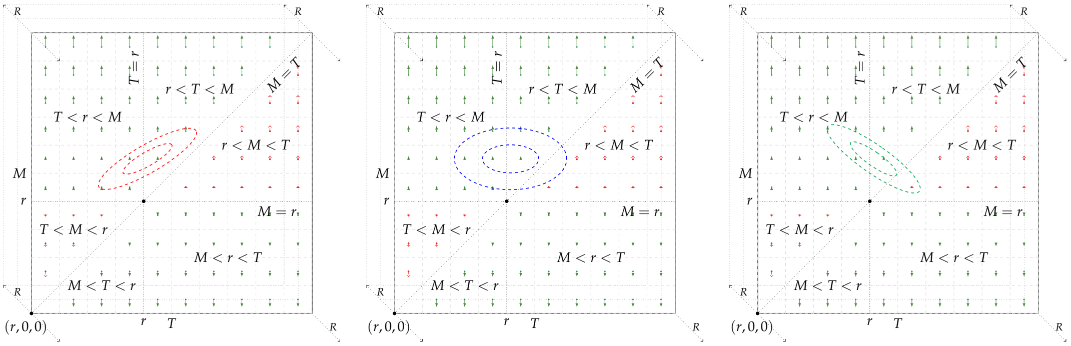

Figure 17.4: A sketch of the distribution of estimates density from Figure 16.1 for the case where the market maker M is biased more than the trader T w.r.t $R = r$. While having such a superior model, being correlated with the market poses no problem to profit, and decreasing the correlation will commonly hurt the performance.

The situation is depicted in Figure 17.4. We can see that in this setting, decreasing $\text{Corr}[T, M|R]$ can easily work in a directly counterproductive fashion by consistently decreasing the profits.

Importantly, however, it should be noted that with a model that is superior by either means, i.e. where $err(\mathsf{t}) < err(\mathsf{m})$, there is no need in trying to decrease $\text{Corr}[T, M|R]$ to make profits, since such a model t needs no help with that to begin with. This can be conveniently detected in advance by measuring $err(\mathsf{t})$ and trading the model with standard investment strategies (Section 2.7) instead.

17.4 THE PROBLEM WITH KELLY

The scenarios we have demonstrated so far operated with the simple uniform investment strategy (Section 2.7), which allowed us to generate profits through decorrelation (Definition 17.0.1), even with models of inferior accuracy w.r.t. the market. As indicated in Section 16.1, let us now explain why this cannot be done with the plain Kelly investment strategy (Section 2.7.3).

We have shown that the growth of wealth W_G with the Kelly strategy is directly equal to difference between the KL-divergence of the market maker from the true distribution $D_{KL}(M||R)$ and the trader from the true distribution $D_{KL}(T||R)$, respectively, in Equation 16.6. It follows that a plain Kelly trader does not care about the particular relationships between R, M, T which are essential to profitability (Section 16.2), since the only thing that matters is how close are our estimates to the true distribution as compared to the bookmaker. The two principally different views of the market distribution P_Ω properties are depicted in Figure 17.5.

Consequently, it thus does not matter whether an individual opportunity seems to have a positive or negative expected profit, since an optimal Kelly trader will bet an

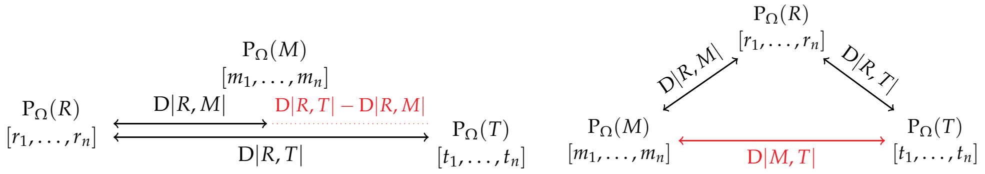

Figure 17.5: The difference between a typical statistical treatment of quality of the estimators, where the relationship between M and T is considered merely in the terms of their distances to R (left), and the proposed scenario, where we explicitly consider their mutual statistical relationship (right).

amount derived merely from the expected growth corresponding to the information advantage w.r.t. R. Given the knowledge of the fair price distribution $P_\Omega(R)$, this provably leads to more wealth than any other strategy. It is then tempting to believe that the Kelly solution then provides an upper-bound to the amount of profit that can be made in any scenario. Nevertheless, this is not true in the real world setting where we lack the knowledge of the true probabilities $P_\Omega(R)$, as we demonstrate on the following example.

Example 17.4.1. Consider a simple scenario of Kelly betting on a match m with two equally probable exclusive outcomes $\{home, away\}$, and the two corresponding opportunities $\omega_h^\beta, \omega_a^\alpha$ priced by the bookmaker m and the trader t equally as follows

$$\mathsf{m}(\omega_t^\beta) = \begin{cases} 0.3 & \text{on } t\text{=home} \\ 0.7 & \text{on } t\text{=away} \end{cases} \qquad \mathsf{t}(\omega_t^\alpha) = \begin{cases} 0.3 & \text{on } t\text{=home} \\ 0.7 & \text{on } t\text{=away} \end{cases} \tag{17.4}$$

Following the derivation from Section 16.1, the optimal vector of fractions $\boldsymbol{f}$ to bet on the outcomes (back the selections) would be

$$f = \begin{cases} 0.3 & \text{on } t\text{=home} \\ 0.7 & \text{on } t\text{=away} \end{cases} \tag{17.5}$$

Since the two estimates of m and t coincide, there is clearly no information advantage of the trader and consequently zero profit to be made with Kelly. And the situation is the same for any other strategy, too, since the expected profitability of the opportunities, following the expected profit definition from Equation 2.10, from the perspective of the trader is simply zero

$$\mathbb{E}_{\mathsf{t}}\left[\rho_h^\alpha\right] = \frac{0.3}{0.3} - 1 = 0 \tag{17.6}$$

$$\mathbb{E}_{\mathsf{t}}\left[\rho_a^\alpha\right] = \frac{0.7}{0.7} - 1 = 0 \tag{17.7}$$

Now consider altering the scenario by decorrelating the estimates of the trader t as follows

$$\mathsf{m}(\omega_t) = \begin{cases} 0.3 & \text{on } t\text{=home} \\ 0.7 & \text{on } t\text{=away} \end{cases} \qquad \mathsf{t}(\omega_t) = \begin{cases} 0.7 & \text{on } t\text{=home} \\ 0.3 & \text{on } t\text{=away} \end{cases} \tag{17.8}$$

Despite switching the estimates, we can clearly see that both the bookmaker m and the trader t are still equally distanced from the fair price r (by the means of D_{KL} as well as any other possible metric), leading again to no information advantage and, as expected (Section 16.1), to zero actual growth of wealth (Section 2.7.3):

$$\begin{aligned} W_{\mathsf{G}} &= \frac{1}{t} \cdot \log\left(\frac{W_t}{W_0}\right) \\ &= \sum_i r_i \cdot \log\left(\frac{f_i}{m_i}\right) \\ &= 0.5 \cdot \log\left(\frac{0.7}{0.3}\right) + 0.5 \cdot \log\left(\frac{0.3}{0.7}\right) \\ &= 0 \end{aligned} \tag{17.9}$$

Nevertheless, the essential profitability of the opportunities from the perspective of the trader is now

$$\mathbb{E}_{\mathsf{t}}[\rho_h^\alpha] = \frac{0.7}{0.3} - 1 = 1.33 \tag{17.10}$$

$$\mathbb{E}_{\mathsf{t}}[\rho_a^\alpha] = \frac{0.3}{0.7} - 1 = -0.57 \tag{17.11}$$

and betting uniformly (Section 2.7.1) some unit on the first, correctly recognized (Section 16.2), opportunity ω_h, the trader would make a consistent unit profit of

$$\mathbb{E}_{\mathsf{r}}[\rho_h^\alpha] = \frac{r_h}{\mathsf{m}(\omega_h^\alpha)} - 1 = \frac{0.5}{0.3} - 1 = 0.66 \tag{17.12}$$

despite it being different from his/her estimated $\rho_h^\alpha = 1.33$.

The inability of Kelly to make profit in such profitable scenarios follows directly from its growth-based view of optimal investments (Section 2.7.3). [4] While maximizing the growth W_{G}, less than optimal investments are just as harmful as over-investment,

[4] Note that the situation in this example remains the same even when allowing Kelly to keep part of the bankroll aside in the cash option (Section 2.7). This would introduce an infinite subspace of other optimal fractional solutions, all of which however, by definition (Section 2.7.3), lead to the exact same growth values (i.e. zero true growth for any optimal Kelly solution here).

despite the latter leading to definite ruin while the former only leads to sub-optimal growth. To distinguish between the two arguably different types of risk, various modifications of the Kelly criterion have been proposed to reflect the natural preference for avoiding the ruin at the cost of sub-optimal growth. By giving up on the expected (theoretical) growth optimality (i.e., w.r.t. $\mathbb{E}_t$), we might then actually achieve better *true* growth (i.e., w.r.t. $\mathbb{E}_r$) in many real world settings, as we demonstrate below.

17.4.1 *Fractional Kelly*

Perhaps the most common remedy to mitigate the risk stemming from the erroneous estimates is fractional Kelly (Section 2.7.3). Let us demonstrate the effect of this risk management practice on the introduced setting from Example 17.4.1 as follows.

Example 17.4.2. Being aware of the uncertainty in her estimates, the Kelly trader now decreases the optimal fraction by one-half, popularly referred to as "half-Kelly" betting. Considering the first scenario of coincidental estimates from Equation 17.4, the invested fractions are now thus decreased by half as

$$f = \begin{cases} 0.15 & \text{on } t\text{=home} \\ 0.35 & \text{on } t\text{=away} \end{cases} \tag{17.13}$$

However, the growth of wealth, accounting for the half of it being held separately, stays inert at zero since

$$\begin{aligned} W_G &= \frac{1}{t} \cdot \log\left(\frac{W_t}{W_0}\right) \\ &= \sum_i r_i \cdot \log\left(0.5 + \frac{f_i}{m_i}\right) \\ &= 0.5 \cdot \log\left(0.5 + \frac{0.15}{0.3}\right) + 0.5 \cdot \log\left(0.5 + \frac{0.35}{0.7}\right) \\ &= 0 \end{aligned} \tag{17.14}$$

In words, the trader is still under-betting the the first opportunity ω_h^α which is profitable, while putting a larger amount on the second opportunity ω_a^α, which is loss-making. Decreasing the fractions then cannot change the simple fact that the opportunity is not recognized correctly by the estimator t w.r.t. essential profitability (Definition 16.2.1).

Consider now the latter scenario of the decorrelated estimates laid out in Equation 17.8. The invested fractions are again cut by half as

$$f = \begin{cases} 0.35 & \text{on } t\text{=home} \\ 0.15 & \text{on } t\text{=away} \end{cases} \tag{17.15}$$

While the information advantage and all the properties of the estimators stay the same as in Equation 17.8, the growth of wealth, accounting for the half of it being held separately, is now positive, particularly

$$\begin{aligned} W_G &= \frac{1}{t} \cdot \log\left(\frac{W_t}{W_0}\right) \\ &= \sum_i r_i \cdot \log\left(0.5 + \frac{f_i}{m_i}\right) \\ &= 0.5 \cdot \log\left(0.5 + \frac{0.35}{0.3}\right) + 0.5 \cdot \log\left(0.5 + \frac{0.15}{0.7}\right) \\ &= 0.087 \end{aligned} \tag{17.16}$$

In words, decreasing the amount invested into the first opportunity ω_h^α from 0.7 down to 0.3, which is now below the optimal $f_\alpha^* = r_h^\alpha = 0.5$ for the full Kelly fraction, removes the overbetting problem, while keeping the second fraction on the loss-making opportunity ω_a^α comparably low.

While the full Kelly forms an interesting corner case, oblivious to the correlation of the estimator w.r.t. the market, we have demonstrated that the decorrelation concept has a positive impact on the commonly used fractional Kelly modification, where decreasing the correlation $\text{Corr}[M, T|R]$ consistently increases profits given that other properties of the distribution stay the same. Similarly, it also improves profitability of other common portfolio optimization strategies, such as the MPT, which we demonstrate practically in experiments (Section 19).

17.5 TOWARDS DECORRELATED ESTIMATORS

So far, we have analysed profitability of the price estimators t w.r.t. various market distributions P_Ω and investment strategies s to derive the desired decorrelation property. Nevertheless, we have not yet discussed how to actually create such profitable estimators t. As outlined in Section 2.5, the common way to create a fundamental price estimator e is by optimizing its fit to historical data $\mathcal{D}$ via minimization of some error measure $err(\mathsf{e})$. The standard desideratum is then to predict the fair prices R within P_Ω as closely as possible.[5]

[5] while keeping the model complexity reasonably low for good generalization.

However, we now know that being accurate is not the only possible desideratum of the estimator, and that for the subsequent trading it might be even more important to minimize the conditional correlation with the market $\mathrm{Corr}[M, T|R]$. Note that an optimal trade-off between these properties of an estimator will naturally depend on the subsequent investment strategy being used (Section 2.7). For instance, we have shown that in the corner case of the full Kelly investor, minimizing the correlation has no impact on the profits at all (Section 17.4) and to maximize the Kelly growth (Section 2.7.3) it is sufficient to resort to the plain cross-entropy error (Section 2.5), minimizing the KL-distance from the fair price distribution (Section 16.1). Nevertheless for the uniform investment strategy (Section 17) as well as other practical strategies, minimizing the covariance will generally improve profitability (Section 17.4.1).

Without discussing the optimal trade-off, we can roughly conclude that it is generally advisable to strive for an unbiased estimator with low variance and low covariance with the market. One can then alter the standard machine learning objectives to reflect these new desiderata, and validate the optimal setting experimentally. The scope of the corresponding error measure would thus include not only the standard estimates T and ground truth values R, but also the market estimates M, and could then look like

$$err^*(R, M, T) = Bias[T|R] + \mathrm{Var}[T|R] + \mathrm{Cov}[T, M|R] \tag{17.17}$$

Recall the purpose of the MSE measure, which is to minimize the squared distance from the fair price as

$$MSE_{\Omega}(R, T) = \mathbb{E}\left[(T - R)^2\right] = \frac{1}{|\Omega|} \sum_{\omega_i \in \Omega} (t_i - r_i)^2 \tag{17.18}$$

which can be thought of as trying to jointly minimize the bias and variance of the estimator (Section 2.5). A straightforward approach to reflect the outlined desiderata is to modify the existing MSE measure with an extra term to also penalize the covariance between M and T as

$$MSE^*_{\Omega}(R, M, T) = \frac{1}{|\Omega|} \sum_{\omega_i \in \Omega} (t_i - r_i)^2 + \gamma \cdot (t_i - r_i)(m_i - r_i) \tag{17.19}$$

where the extra "decorrelation term" is weighted by a tunable hyperparameter $\gamma > 0$. The purpose of γ is then to adjust the trade-off between the decorrelation term and the standard MSE, since the two normally represent opposing objectives, as the market maker tends to be very close to the fair price.

We note that altering the MSE might seem counter-intuitive from the perspective of a perfect estimator, corresponding to the minimal MSE, where the additional term will only hurt its performance by pushing it away, not only from the market price, but from the fair price, too, since $argmin[MSE] \neq argmin[MSE^*]$. However, recall that the decorrelation only makes sense when one is not able to train such a superior model (Section 17.3). In those common cases, the model generally occupies some wider area of the error landscape around the MSE minima, where the additional term is meant to navigate away from the correlated regions (Figure 16.1), which may be equally distanced from the MSE minimum, but will yield inferior profits.

While we do not attempt to argue about optimality of the proposed MSE^* metric, and we acknowledge that there are likely more appropriate measures to maximize the model profitability, we will demonstrate that this simple MSE-extension already works well enough in practice to proof viability of the decorrelation concept.

18

MARKET TAKER'S MODEL

We have already seen in Section 12.3 that bookmakers pose a tough challenge. Moreover, we discovered that the predictions of various score-based models are very similar (Section 7.1). As we aim to explore the accuracy-decorrelation tradeoff, we opt for a more complex model, allowing us to produce predictions with different levels of decorrelation and accuracy. To further enhance the diversity of the predictions we stray from using solely the score-derived features and opt for using much more detailed (player-level) statistics gather during each game. We also describe the details of implementing the MPT (Section 2.7.2) in Section 18.4.

18.1 DATA FEATURES

The information we use for predicting the outcome of a match combines data relating to the home team and those pertaining to the visiting team. For each of the two, we aggregate various quantitative measures of the team's performance in all of its preceding matches since the beginning of the season in which prediction takes place. [1] The entire range of these measures is described in the Appendix a. Current seasons are commonly deemed the most relevant time-windows for player and team performance prediction. The seasonal aggregation is conducted as follows. All variables depending on the match duration are divided by the duration in minutes, and for the seasonal aggregate, we consider the average of these per-minute values. Such variables are marked as "per-minute" in the Appendix. For the remaining variables, the median value is considered instead.

The inputs $d \in D$ to the predictive model t are tuples of real-valued features constructed out of the said season-aggregated data. Some of the variables in the latter pertain to individual players and others relate to the whole team. Consequently, we distinguish two levels of feature-based description. In the fine-grained *player-level*, we collect all player-related variables as individual features, whereas the *team-level* description involves only the team-level variables as features.

[1] A few initial games of the season are thus not included among training instances and serve only to collect the statistics. We will quantify this arrangement for a particular testing data set in Section 19.2.

Meta-parameter	Standard (team-level)	Convolutional (player-level)
Architecture	D64-D32-D16-D1	C1-D64-D16-D1
Activations	tanh	tanh
Dropout	0.2	0.2
L2 regularization	0.0001	0.001

Table 18.1: The architecture and meta-parameters of the neural predictive models considered.

Besides the historical track data considered above, the bookmaker's odds assigned to a match represent another piece of information potentially relevant to the prediction of its outcome. While the odds clearly present a very informative feature, their incorporation in a model naturally increases the undesirable correlation with the bookmaker (Section 18.3). Whether to omit or include the odds as a feature thus remains questionable and so we further consider both the options in the experiments.

18.2 neural model

We explored two variants of a neural network. The first has a standard (deep) feed-forward architecture with 4 dense layers, while the second one uses a *convolutional layer* [73] followed by 3 dense layers. Table 18.1 describes the architectures and the relevant meta-parameters of the two neural models.

The standard feed-forward network is intended for the team-level feature data. The convolutional network is specifically designed for the player-level data to deal with the large number of features involved. The principle of its operation, inspired by well-known applications of convolutional networks for visual data processing, is explained through Figure 18.1. Intuitively, the convolutional layer may be perceived as a bridge from player-level variables to a team-level representation. However, whereas team-level variables already present in the data are simple sums or averages over all team's players, the convolution layer provides the flexibility to form a more complex aggregation pattern, which itself is learned from the training data.

18.3 model decorrelation

In sports betting, one cannot observe the true probability of an opportunity, even retrospectively after the match. Instead, the label $r_i \in \{0, 1\}$ provided with each learning sample ω_i reflects merely the binary outcome realization endowed by the underlying probability. This poses a problem to the decorrelation term from Equation 17.19 which, using the binary realizations instead of the actual probability values, would degenerate

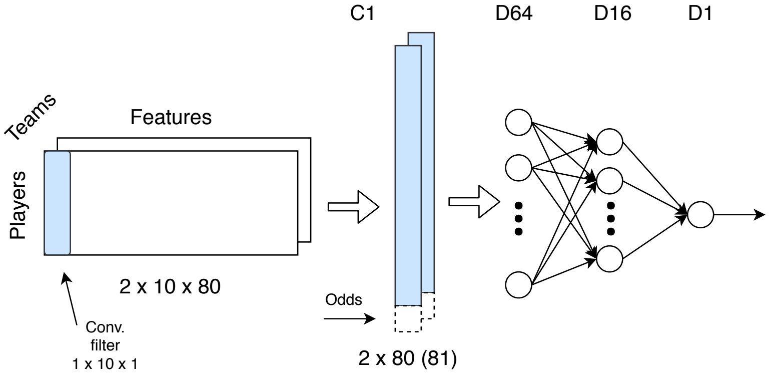

Figure 18.1: The convolutional neural network for player-level data. The input to the network are two matrices (one for the home team, one for the visitors), with players in rows and all player-level features in columns. The rows are sorted by the time-in-play of the corresponding players, and only the top 10 players w.r.t. this factor are included. The convolution layer is defined by a vector of 10 tunable real weights. The output of the layer is a vector where each component is the dot product of the former vector with one of the matrix columns. The vector may be viewed as a filter sliding horizontally on the the first input matrix, and then on the second.

severely. Particularly, the decorrelated cases of $m_i < r_i < t_i$ and $t_i < r_i < m_i$ would be impossible to achieve with the binary values of $r_i \in \{0,1\}$. To better accommodate the proposed MSE^* loss into this practical binary betting setting, we thus slightly alter the loss to directly decorrelate M and T instead as

$$MSE^*_\Omega(R,M,T) = \mathbb{E}[(T-R)^2 - \gamma \cdot (T-M)^2] = \frac{1}{|\Omega|} \sum_{\omega_i \in \Omega} (t_i - r_i)^2 - \gamma \cdot (t_i - m_i)^2 \quad (18.1)$$

where the m_i are the bookmaker's probability estimates induced from the published odds o_i. We note that this is different from the direct penalization of the covariance of the residuals in Equation 17.19. evertheless, the motivation is that a similar effect should be achieved through the integration with the remaining MSE term, which forces the estimates t_i to be unbiased w.r.t. r_i. Hence simply penalizing model estimates that are too close to the market price should also decrease the partial covariance between the two.

18.4 BETTING STRATEGY

Firstly, we evaluated the models by betting with the basic uniform investment strategy (Section 2.7.1), for which we derived the theoretical reasoning in Section 17. We further refer to this simple strategy as unif. Additionally, we also evaluated the classic portfolio optimization technique of Markowitz (Section 2.7.2). For this strategy we needed to calculate the expected returns of the opportunities. Following the probabilistic setting detailed in Section 2.6, the expected profit can be defined as

$$\mathbb{E}_R[w_i] = (\frac{r_i}{m_i} - 1) \cdot f_i. \tag{18.2}$$

We further needed co calculate the covariances of the expected returns (Section 2.7.2). Assuming that there is no correlation between the traded selections, corresponding to the underlying games played within a single round, it is sufficient to consider only the variances of the individual independent opportunities[2] instead of the whole covariance matrix Σ. Following from the underlying Bernoulli distribution, the variance of each opportunity can then be defined as

$$\mathrm{Var}_R[w_i] = \mathbb{E}[w_i^2] - \mathbb{E}[w_i]^2 = (1 - r_i)r_i f_i^2 \frac{1}{{m_i}^2} \tag{18.3}$$

Naturally, we used the model estimates t_i instead of the true (unknown) values r_i in the actual calculations of both $\mathbb{E}_T[w_i]$ and $\mathrm{Var}_T[w_i]$. We then chose the particular portfolio $\boldsymbol{f}$ following the Sharpe ratio criterion (Section 2.7.2), and used the algorithm of sequential quadratic programming [15] to identify its unique maximizer. We further refer to this strategy as sharpe.

[2] This is a somewhat simplifying assumption based on excluding the option to bet on both the exclusive game outcomes simultaneously.

19

EXPERIMENTS

To put the concept of decorrelation to test, we conducted experiments both on generated and real-worl data. The simulations are a necessary part of the experiments as they allow us emphasize the effects of the possible setups (in our case the $\mathrm{Corr}[T,R]$ and $\mathrm{Corr}[T,M]$). The experiments or real-world data are invaluable as they ground the parameters beyond our control in reality.

19.1 SIMULATIONS

To further demonstrate the proposed decorrelation concept (Definition 17.0.1), we conduct an auxiliary experiment requiring simulated data. Here we simulated the ground truth R as well as both of the estimates T, M with various levels of their correlation, and measured the profits made by the sharpe and unif strategies for these different levels.

More precisely, we sampled triples R, T, M from a multivariate Beta distribution. The distribution is parameterized with the marginal means and variances of the three variables and their pair-wise correlations. The mean of each of the three variables was set to 0.5, reflecting the mean probability of the complementary binary outcomes. The variance of M was determined as 0.054 from real bookmaker's data (Section 19.2), and T's variance copies this value. The variance of R was set to 0.08. These values set an upper-bound for *Accuracy*. When R is sampled with mean 0.5 and variance 0.08, then with 0.75 probability the event $(R > 0.5)$ predicts correctly the outcome of a Bernoulli trial parameterized with R.

We let the correlations $\mathrm{Corr}[T,R]$ and $\mathrm{Corr}[T,M]$ range over the values $\{0.85, 0.90, 0.95\}$. These represent the independent variables of the analysis, acting as factors in Table 19.1. $\mathrm{Corr}[R,M]$ was set to 0.9, so we could explore all possible orderings of $\mathrm{Corr}[R,T]$ and $\mathrm{Corr}[R,M]$ w.r.t. $\mathrm{Corr}[]$. For each setting of $\mathrm{Corr}[T,R]$ and $\mathrm{Corr}[T,M]$, we drew r_i, t_i, m_i ($i = 1, 2, \ldots, n = 15$) samples, to simulate one round of betting. Then we set the odds $o_i = 1/m_i$ (the bookmaker's margin being immaterial here) for $1 \leq i \leq n$, and determined the bets $f_1, f_2, \ldots, f_n$ from $o_1, o_2, \ldots, o_n$ and $t_1, t_2, \ldots, t_n$ using the sharpe and unif strategies. Finally, the match outcomes were established by a Bernoulli trial

$\text{Corr}[T, R]$	$\text{Corr}[T, M]$	ρ_{sharpe}	ρ_{unif}	Acc.	Cons.	Upset	Missed	Spotted
0.85	0.85	0.12	0.05	70.05	61.96	20.40	9.55	8.09
	0.90	0.07	0.02	70.04	63.53	22.10	7.85	6.51
	0.95	-0.02	-0.04	70.06	65.65	24.23	5.71	4.41
0.90	0.85	0.19	0.11	71.46	62.62	19.72	8.82	8.84
	0.90	0.15	0.09	71.44	64.20	21.36	7.20	7.24
	0.95	0.10	0.05	71.53	66.51	23.41	5.06	5.02
0.95	0.85	0.26	0.16	72.85	63.29	18.99	8.16	9.56
	0.90	0.23	0.15	72.90	64.96	20.61	6.49	7.95
	0.95	0.21	0.14	72.97	67.23	22.69	4.34	5.74

Table 19.1: Returns ρ_{sharpe} of the sharpe and ρ_{unif} of the unif strategies w.r.t. to the correlations of the (estimated) probabilities. Accuracy denotes the % of correct outcome predictions by the bettor (predict win if $t_i > 0.5$). The four last columns break down the proportions (in %) of different combinations of predictions by t (bettor) and m (bookmaker): *Consensus* (both right), *Upset* (both wrong), *Missed* (bettor wrong, bookmaker right), *Spotted* (bettor right, bookmaker wrong).

for each of the $r_1, r_2, \ldots, r_n$. This procedure was repeated 30 000 times (rounds). [1] With these inputs, we calculated the ρ of sharpe and unif strategy.

Table 19.1 then shows the returns as well as the accuracy of the bettor's outcome predictions (call win if $T > 0.5$), and the percentual breakdown of 4 possible combinations of bettor's and bookmaker's predictions. The accuracies, as well as the four latter proportions, are also averaged over all bets in all simulated rounds.

Besides the unsurprising observation that bettor's prediction accuracy grows with $\text{Corr}[T, R]$, the results show that profits indeed decay systematically as the bettor's and bookmaker's predictions become more correlated (increasing $\text{Corr}[T, M]$ decreases profit). An instructive observation is that the proportion of *spotted* opportunities is in all cases higher when the bookmaker's and bettor's predictions are less correlated. Furthermore, we observe the effect of market taker's advantage (Section 16.3) in instances where $\text{Corr}[T, R] = \text{Corr}[M, R]$. Moreover, we can see that the betting strategy is another independent factor strongly influencing the profit, with the sharpe strategy being superior to the unif strategy. Besides the achieved return being higher, the variance of the return between rounds of the sharpe strategy (≈ 0.1) was half of the variance of the unif stategy (≈ 0.2).

[1] Note that this is not the same (for W_{sharpe}) as setting $n = 15 \cdot 30000$ without repeating the procedure, as the full unit budget is supposed to be spent in each round.

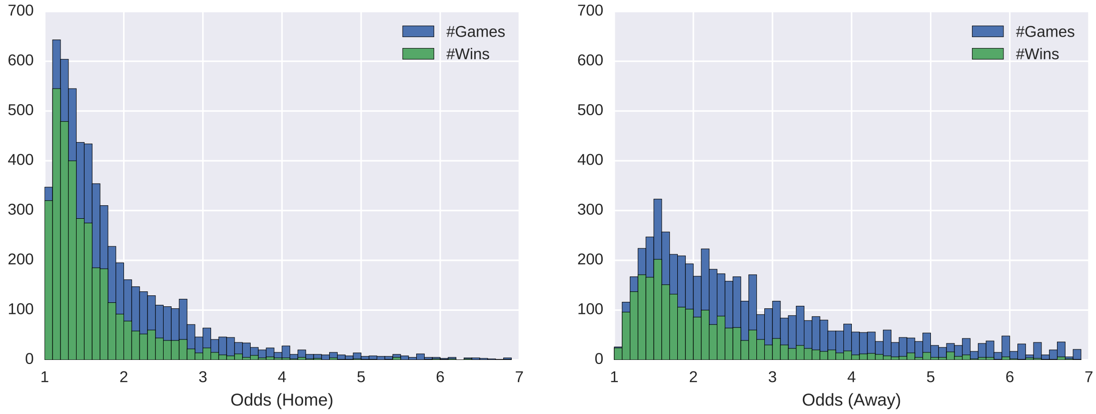

Figure 19.1: Distribution of all games (blue) with respective proportions of wins (green) w.r.t odds set by the bookmaker from home (left) and away (right) team perspectives. Clearly, the home team is generally favored by the bookmaker, with the true proportions roughly following the inverse of odds.

19.2 DATA

We retrieved the official box score data from the National Basketball Association (NBA) from seasons 2000 to 2014. The gathered data provide game summaries; namely, player-level and team-level statistics such as the number of shots or number of steals per game are recorded. The detailed description of the available kinds of information can be found in Appendix a. Games with incomplete statistics were removed, and thus the number of games differs slightly between seasons; on average, 985 games per year were included. 10 initial games of each team in each season were not included as training instances as they only served for the initial calculation of seasonal aggregates (c.f. Section 18.1). There are 30 teams in the NBA, so one league round consists of $n = 15$ games.

For betting odds, we used the *Pinnacle*[2] closing odds for seasons 2010–2014. For earlier seasons, we had to collect odds data from multiple different bookmakers. Figure 19.1 shows histograms of odds distribution for the home and away teams and their winnings, respectively. The histograms reveal that in most matches the home team is the favorite in bookmaker's eyes. This comes as no surprise due to the home court advantage (home teams win in about 60 % of games). Both histograms exhibit long-tailed distributions, as expected given that odds correspond to inverse probabilities, which roughly follow the true proportions of the respective winnings.

Figure 19.2 shows the seasonal averages of the bookmaker's margin ϵ, displaying the artifact caused by different sources of odds information prior and post 2010. This artifact does not confound the experimental questions below, except for causing higher

[2] https://www.pinnacle.com/

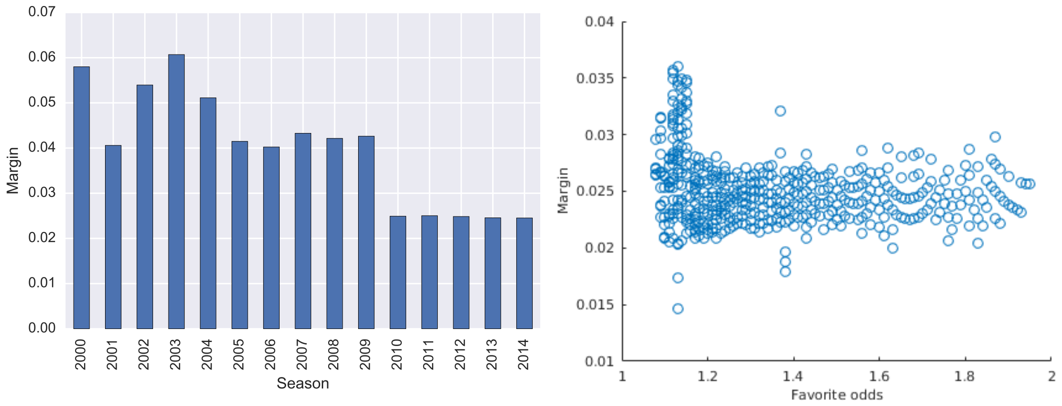

Figure 19.2: Evolution of margin over the seasons (left), showing drop for seasons 2010-2014 where Pinnacle was the only source, and its dependency on bookmaker's odds for the favorite of each game (right), displaying interesting patterns with rapid growth towards the clear favorite case (minimal odds).

profits in late seasons due to the systematically smaller margins. To get a better insight into the bookmaker's margins, we plotted their dependency on odds for the 2010–2014 period. Figure 19.2 indicates a significantly larger margin in the case where there is a clear favorite with high probability of winning (odds close to 1). This is due to an asymmetry in bookmaker's odds: while there are several occasions with the favorite's odds around 1.1 implying win-probability around 91%, odds around 11 representing the complementary probability 9% are extremely rare. This asymmetry is increasing with favorite's odds approaching 1.0.

19.3 EXPERIMENTAL PROTOCOL

The central experimental questions are: how accurate the learned predictors of match outcomes are, how profitable the betting strategies using the predictions are, and how the profitability is related to the correlation between the bookmaker's and bettor's models. The selection portfolios for the betting strategies were then formed repeatedly from the batches of 15 matches played in each round of the NBA league, which were available simultaneously on the betting market. The profit from each such round was then evaluated independently, i.e. the same unit of budget was assumed for staking in each round. Note that this repeated single-period portfolio optimization setting (Section 2.7.2) is different from the growth-based view on portfolio optimization in time, as assumed by Kelly (Section 2.7.3).

Training and evaluation of the models and betting strategies followed the natural chronological order of the data w.r.t individual seasons, i.e. only past seasons were ever used for training a model evaluated on the upcoming season. To ensure sufficient training data, the first season to be evaluated was 2006, with a training window made up of seasons 2000–2005, iteratively expanding all the way to evaluation on 2014, trained on the whole preceding range of 2000–2013.

19.4 RESULTS

The number of games with complete statistics available varies slightly with each individual season providing around 1000–1050 matches. The total number of 9093 games from the evaluated seasons 2006–2014 is counted towards the accuracies (% of correctly predicted outcomes) of each model, whose results are displayed in Table 19.2. The accuracy of the bookmakers' model, predicting the team with smaller odds to win, levels over these seasons at 69 ± 2.5%. Generally in terms of accuracy, the bookmakers' model is slightly superior to the neural models.

As expected, we can observe from the results that models utilizing the highly informative odds feature achieve consistently higher accuracies. Similarly, the models that included the bookmakers' odds were anticipated to be more correlated with the bookmaker. This is convincingly confirmed by measurements of Pearson coefficients which stay at 0.87 for the models trained without odds as a feature, and 0.95 for models including them, applying equally to both the player-lever and team-level models.

Table 19.2 also provides important insights on the profit generation. We display two selected betting strategies (sharpe, unif) against a range of considered variants of predictive models. Similarly to the simulation experiment, the superiority of sharpe over unif is still evident. Apart from accuracy, we argued for decorrelation as an important factor for profit, which we here enforce by the means of the altered loss function while varying the trade-off γ between accuracy and decorrelation. We can clearly see that such a trade-off is effectively possible for a wide range of $0.4 \leq \gamma \leq 0.8$ resulting into positive returns over all the models utilizing the sharpe strategy.

In Figure 19.3 we display insight on how the trade-off constant γ influences the distribution of the four betting outcome situations. As expected, increasing the decorrelation results in a desirable increase of *spotted* opportunities, i.e., cases where the model correctly predicted the underdog's victory. If this increase is too abrupt, however, it is outweighed by the parallel increase of *missed* opportunities where the bet on the underdog was wrong.

Revisiting the question as to whether include the odds feature or not, in terms of profit generation the results are inconclusive, with the team-level model performing slightly better with the feature and the player-level model without it.

Next we investigate the effects of confidence-thresholding used to filter the predictions (Section 16.5) before providing them to the betting strategy. By varying the threshold ϕ we can trade off between the confidence of the model and the number of games providing information to the strategy. Results in Table 19.3 are conclusive in that a reasonably low amount of thresholding $\phi \leq 0.2$ in conjunction with the sharpe strategy

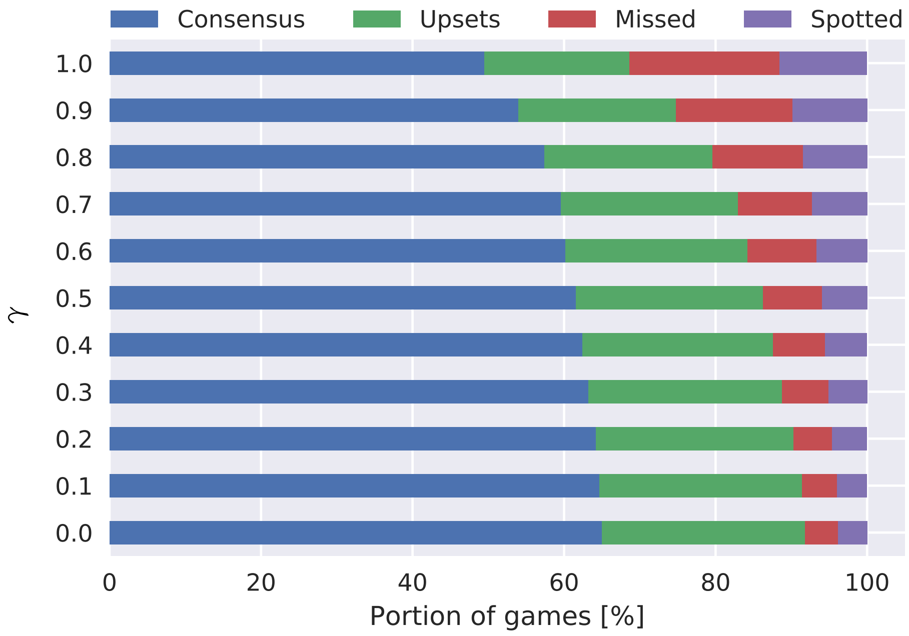

Figure 19.3: The impact of loss-function-term model-decorrelation techniques, as introduced in Section 18.3 and applied on the team-level model, on the distribution of betting opportunity outcomes: *Consensus* (both right), *Upset* (both wrong), *Missed* (bettor wrong, bookmaker right), *Spotted* (bettor right, bookmaker wrong).

		Team-level			*Player-level*		
	γ	ρ_{sharpe}	ρ_{unif}	*Accuracy*	ρ_{sharpe}	ρ_{unif}	*Accuracy*
Without odds	0.0	-0.94 ± 0.12	-4.31 ± 0.17	67.47 ± 0.05	0.38 ± 0.10	-5.12 ± 0.11	67.62 ± 0.03
	0.2	-0.58 ± 0.14	-3.60 ± 0.19	67.39 ± 0.04	1.05 ± 0.12	-3.31 ± 0.13	67.47 ± 0.03
	0.4	0.46 ± 0.15	-1.94 ± 0.20	67.30 ± 0.05	1.74 ± 0.14	-1.73 ± 0.18	67.15 ± 0.10
	0.6	0.86 ± 0.08	-1.68 ± 0.22	66.93 ± 0.06	1.32 ± 0.14	-0.61 ± 0.28	66.19 ± 0.09
	0.8	1.37 ± 0.08	-0.79 ± 0.16	65.94 ± 0.12	1.10 ± 0.29	-0.39 ± 0.22	64.93 ± 0.35
	1.0	-1.06 ± 0.35	-1.32 ± 0.31	61.38 ± 0.19	-1.92 ± 0.81	-2.59 ± 0.57	61.30 ± 0.48
With odds	0.0	0.89 ± 0.10	-2.24 ± 0.21	68.83 ± 0.05	-0.12 ± 0.24	-3.83 ± 0.22	68.80 ± 0.06
	0.2	0.92 ± 0.18	-2.10 ± 0.24	68.71 ± 0.04	0.72 ± 0.13	-2.50 ± 0.14	68.37 ± 0.04
	0.4	1.24 ± 0.12	-1.24 ± 0.22	68.42 ± 0.05	1.49 ± 0.10	-1.30 ± 0.12	67.48 ± 0.10
	0.6	1.44 ± 0.11	-0.64 ± 0.21	67.88 ± 0.06	1.02 ± 0.20	-1.15 ± 0.22	66.55 ± 0.10
	0.8	1.41 ± 0.10	-0.56 ± 0.20	66.64 ± 0.12	1.00 ± 0.35	-0.45 ± 0.28	65.19 ± 0.27
	1.0	-0.37 ± 0.16	-0.74 ± 0.13	62.49 ± 0.12	-1.22 ± 0.51	-2.25 ± 0.30	61.77 ± 0.44

Table 19.2: Averages and standard errors of profits (from 10 runs over seasons 2006–2014) for the two strategies (sharpe, unif) with accuracies of *Player-level* and *Team-level* outcome prediction models (Section 18) across different levels of decorrelation (Section 18.3).

indeed improves profits. Such a low threshold has the effect of filtering out generally those predictions that are indifferent on the winner (estimated probabilities of 0.5 ± 0.2), which was the main motivation for this technique.

		ϕ	ρ_{sharpe}	ρ_{unif}	*Accuracy*	*Bets*	*Stake*
Team-level	*Without odds*	0.0	0.86 ± 0.08	-1.68 ± 0.22	66.93 ± 0.06	9093	602
		0.1	1.61 ± 0.14	-1.37 ± 0.21	70.31 ± 0.07	7370	602
		0.2	1.99 ± 0.25	-1.26 ± 0.21	74.08 ± 0.13	5442	601
		0.3	0.54 ± 0.62	-2.65 ± 0.73	79.64 ± 0.20	2937	577
	With odds	0.0	1.44 ± 0.11	-0.64 ± 0.21	67.88 ± 0.06	9093	602
		0.1	2.18 ± 0.14	-0.13 ± 0.25	70.93 ± 0.06	7538	602
		0.2	1.80 ± 0.24	-0.73 ± 0.29	74.47 ± 0.09	5749	602
		0.3	0.78 ± 0.35	-1.65 ± 0.39	80.26 ± 0.21	3315	584
Player-level	*Without odds*	0.0	1.74 ± 0.14	-1.73 ± 0.18	67.15 ± 0.10	9093	602
		0.1	2.39 ± 0.20	-1.42 ± 0.16	72.01 ± 0.15	6686	602
		0.2	3.24 ± 0.32	-1.18 ± 0.29	77.22 ± 0.19	4228	601
		0.3	-5.41 ± 0.92	-7.76 ± 1.22	84.32 ± 0.48	1841	533
	With odds	0.0	1.49 ± 0.10	-1.30 ± 0.12	67.48 ± 0.10	9093	602
		0.1	2.43 ± 0.16	-0.94 ± 0.24	72.2 ± 0.08	6749	602
		0.2	3.39 ± 0.46	-0.70 ± 0.53	77.41 ± 0.12	4336	600
		0.3	-5.06 ± 1.05	-8.12 ± 0.95	84.35 ± 0.29	1940	545

Table 19.3: Averages and standard errors of profits (from 10 runs over seasons 2006–2014) for the two strategies (sharpe, unif) with accuracies of the *Player-level* ($\gamma = 0.4$) and *Team-level* ($\gamma = 0.6$) prediction models (Section 18) across different levels of confidence thresholding (Section 16.5). *Bets* represent numbers bets placed and *Stake* is the total amount staked.

20

CONCLUSION

The main hypotheses of this study were 1) that correlation of outcome predictions with the bookmaker's predictions is detrimental for the bettor, and that suppressing such correlation will result in models allowing for higher profits, 2) that convolutional neural networks are a suitable model to leverage player-level data for match outcome predictions, and 3) that a successful betting strategy should balance optimally between profit expectation and profit variance.

The first hypothesis was clearly confirmed in simulated experiments and also supported by extensive real-data experiments. In the former, for each level of constant accuracy (correlation of model and ground truth), increasing the correlation between the model and the bookmaker consistently decreased the profit in all settings. In the latter, models trained with the proposed decorrelation loss achieved higher profits despite having lower accuracies than models with higher correlation, in all settings up to a reasonable level of the decorrelation-accuracy trade-off.

Regarding the second hypothesis, the convolutional network achieved generally higher accuracies and profits than the rest of the models in the settings excluding bookmaker's odds from features. This can evidently be ascribed to its ability to digest the full matrix of players and their performance statistics through a flexible (learnable) pattern of aggregation, as opposed to just replicating the bookmaker's estimate from input.

As for the third hypothesis, the portfolio-optimization sharpe strategy consistently dominated the standard unif strategy both in simulations and in experiments on real-world data. Additionally, we proposed confidence-thresholding as an enhancement to the strategy when used in conjunction with models utilizing logistic sigmoid output. This technique effectively removes very uncertain predictions from the strategy, leading to additional increases in profit.

Part V

CONCLUSIONS

21

CONCLUSIONS

In this thesis, we delved into the intricate task of modeling the probabilities of winning a sports event with the aim of achieving profitability in the market.

In Part ii, we conducted an extensive experimental review of existing score-based models, categorizing them into two distinct groups: rating systems and statistical models. Our reimplementation and comparison of nine previously published models on the largest publicly available dataset revealed remarkably similar performance. Further scrutiny of the models' predictions unveiled shared similarities, especially within the same category. This comprehensive review addressed a longstanding gap in the domain, as previous comparisons lacked robustness due to limited data or were altogether absent.

Having established the state-of-the-art performance, we designed and developed our own models in Part iii. We experimented with feature-based classification and regression models as well as classifier leveraging the relational nature of the data (LRNNs). While the feature-based regression model was dominated by its' classification counterpart, the LRNNs proved to be competitive in terms of performance. However, the model was too computationally demanding and, therefore, not comparable with the state-of-the-art models on large-scale data. The feature-based classifier outperformed the state-of-the-art models across all examined metrics by a significant margin. We demonstrated how easily a simpler model can be incorporated into this model and how carefully engineered score-derived features can further enhance its' predictive capabilities. Despite the significant improvement over the state-of-the-art, the predictive power of bookmakers' odds remained far out of reach. This result called for the use of more complex models or a different approach for beating the market than relying on universally more accurate predictions.

Finally, in Part iv we focused on trading our predictions on the markets. We designed and implemented a neural model for the NBA competition, where detailed player-level data are gathered for each game. We outlined the requirements for profitability and formalized the often-overlooked market taker's advantage. Introducing the concept of decorrelation, we illustrated that profits on the market could be achieved even with a model inferior in standard accuracy-based metrics, provided it generated decorrelated predictions. Real-world data testing validated the crucial role of decorrelation in beating the market.

FUTURE WORK While we have demonstrated that score-based approaches significantly lag behind bookmakers, we believe these models could be utilized in other domains or lower-tier leagues where more granular data is not collected. For high-profile events, the exploration of player-level data appears promising. Our experiments showed that such data adds value compared to using only team-level data.

The concept of decorrelation, introduced in this study, opens up new possibilities for future research. We employed a modified mean squared error loss function during model fitting to encourage dissimilar predictions from the market. However, exploring other modified loss functions may achieve an even better tradeoff between accuracy and decorrelation.

In our experiments, we combined the Modern Portfolio Theory with confidence thresholding to maximize returns while managing risk. Nevertheless, other methods for selecting and evaluating opportunities warrant exploration.

Part VI

APPENDIX

a

NBA STATISTICS

Below is the list of player and team performance data we used for constructing features for the model desctibed in Section 18. The grouping of variables and the acronyms shown match the source of the data `http://stats.nba.com`.

BASIC STATISTICS

- AST: Number of assists. An assist occurs when a player completes a pass to a teammate that directly leads to a field goal. *(per minute)*
- BLK: Number of blocks. A block occurs when an offensive player attempts a shot, and the defense player tips the ball, blocking their chance to score. *(per minute)*
- DREB: Number of rebounds a player or team has collected while they were on defense. *(per minute)*
- FG_PCT: Percentage of field goals that a player makes. The formula to determine field goal percentage is: Field Goals Made/Field Goals Attempted. *(per minute)*
- FG3_PCT: Percentage of 3 point field goals that a player or team has made. *(per minute)*
- FG3A: Number of 3 point field goals that a player or team has attempted. *(per minute)*
- FG3M: Number of 3 point field goals that a player or team has made. *(per minute)*
- FGA: Number of field goals that a player or team has attempted. This includes both 2 pointers and 3 pointers. *(per minute)*
- FGM: Number of field goals that a player or team has made. This includes both 2 pointers and 3 pointers. *(per minute)*
- FT_PCT: Percentage of free throws that a player or team has made.
- FTA : Number of free throws that a player or team has taken. *(per minute)*
- FTM: Number of free throws that a player or team has successfully made. *(per minute)*
- MIN: Number of minutes a player or team has played.
- OREB: Number of rebounds a player or team has collected while they were on offense. *(per minute)*

- PF: Number of fouls that a player or team has committed. *(per minute)*
- PLUS_MINUS: Point differential of the score for a player while on the court. For a team, it is how much they are winning or losing by. *(per minute)*
- PTS: Number of points a player or team has scored. A point is scored when a player makes a basket. *(per minute)*
- REB: Number of rebounds: a rebound occurs when a player recovers the ball after a missed shot. *(per minute)*
- STL: Number of steals: a steal occurs when a defensive player takes the ball from a player on offense, causing a turnover. *(per minute)*
- TO: Number of turnovers: a turnover occurs when the team on offense loses the ball to the defense. *(per minute)*

ADVANCED STATISTICS

- AST_PCT: Assist Percentage - % of teammate's field goals that the player assisted.
- ST_RATIO: Assist Ratio - number of assists a player or team averages per 100 of their own possessions.
- AST_TOV: Number of assists a player has for every turnover that player commits.
- DEF_RATING: Number of points allowed per 100 possessions by a team. For a player, it is the number of points per 100 possessions that the team allows while that individual player is on the court.
- DREB_PCT: The percentage of defensive rebounds a player or team obtains while on the court.
- EFG_PCT: Effective Field Goal Percentage is a field goal percentage that is adjusted for made 3 pointers being 1.5 times more valuable than a 2 point shot.
- NET_RATING: Net Rating is the difference in a player or team's Offensive and Defensive Rating. The formula for this is: Offensive Rating-Defensive Rating.
- OFF_RATING: The number of points scored per 100 possessions by a team. For a player, it is the number of points per 100 possessions that the team scores while that individual player is on the court.
- OREB_PCT: The percentage of offensive rebounds a player or team obtains while on the court.
- PACE: The number of possessions per 48 minutes for a player or team.
- PIE: An estimate of a player's or team's contributions and impact on a game: the % of game events that the player or team achieved.
- REB_PCT: Percentage of total rebounds a player obtains while on the court.
- TM_TOV_PCT: Turnover Ratio: the number of turnovers a player or team averages per 100 of their own possessions.
- TS_PCT: A shooting percentage that is adjusted to include the value three pointers and free throws.
- USG_PCT: Percentage of a team's offensive possessions that a player uses while on the court.

FOUR FACTORS, AS DESCRIBED BY [68]

- EFG_PCT: Effective Field Goal Percentage is a field goal percentage that is adjusted for made 3 pointers being 1.5 times more valuable than a 2 point shot.
- FTA_RATE: The number of free throws a team shoots in comparison to the number of shots the team attempted. This is a team statistic, measured while the player is on the court. The formula is Free Throws Attempted/Field Goals Attempted. This statistic shows who is good at drawing fouls and getting to the line.
- OPP_EFG_PT: Opponent's Effective Field Goal Percentage is what the team's defense forces their opponent to shoot. Effective Field Goal Percentage is a field goal percentage that is adjusted for made 3 pointers being 1.5 times more valuable than a 2 point shot.
- OPP_FTA_RATE: The number of free throws an opposing player or team shoots in comparison to the number of shots that player or team shoots.
- OPP_OREB_PCT: The opponent's percentage of offensive rebounds a player or team obtains while on the court.
- OPP_TOV_PCT: Opponent's Turnover Ratio is the number of turnovers an opposing team averages per 100 of their own possessions.
- OREB_PCT: The percentage of offensive rebounds a player or team obtains while on the court.
- TM_TOV_PCT: Turnover Ratio is the number of turnovers a player or team averages per 100 of their own possessions.

PLAYER SCORING STATISTICS

- PCT_AST_2PM: % of 2 point field goals made that are assisted by a teammate.
- PCT_AST_3PM: % of 3 point field goals made that are assisted by a teammate.
- PCT_AST_FGM: % of field goals made that are assisted by a teammate.
- PCT_FGA_2PT: % of field goals attempted by a player or team that are 2 pointers.
- PCT_FGA_3PT: % of field goals attempted by a player or team that are 3 pointers.
- PCT_PTS_2PT: % of points scored by a player or team that are 2 pointers.
- PCT_PTS_2PT_MR: % of points scored by a player or team that are 2 point mid-range jump shots. Mid-Range Jump Shots are generally jump shots that occur within the 3 point line, but not near the rim.
- PCT_PTS_3PT: % of points scored by a player or team that are 3 pointers.
- PCT_PTS_FB: % of points scored by a player or team that are scored while on a fast break.
- PCT_PTS_FT: % of points scored by a player or team that are free throws.
- PCT_PTS_OFF_TOV: % of points scored by a player or team that are scored after forcing an opponent's turnover.

- PCT_PTS_PAINT: % of points scored by a player or team that are scored in the paint.
- PCT_UAST_2PM: % of 2 point field goals that are not assisted by a teammate.
- PCT_UAST_3PM : % of 3 point field goals that are not assisted by a teammate.
- PCT_UAST_FGM: % of field goals that are not assisted by a teammate.

USAGE STATISTICS

- PCT_AST: % of team's assists a player contributed.
- PCT_BLK: % of team's blocked field goal attempts a player contributed.
- PCT_BLKA: % of team's blocked field goal attempts a player contributed.
- PCT_DREB: % of team's defensive rebounds a player contributed.
- PCT_FG3A: % of team's 3 point field goals attempted a player contributed.
- PCT_FG3M: % of team's 3 point field goals made a player contributed.
- PCT_FGA: % of team's field goals attempted a player contributed.
- PCT_FGM: % of team's field goals made a player contributed.
- PCT_FTA: % of team's free throws attempted a player contributed.
- PCT_FTM: % of team's free throws made a player contributed.
- PCT_OREB: % of team's offensive rebounds a player contributed.
- PCT_PF: % of team's personal fouls a player contributed.
- PCT_PFD: % of team's personal fouls drawn a player contributed.
- PCT_PTS: % of team's points a player contributed.
- PCT_REB: % of team's rebounds a player contributed.
- PCT_STL: % of team's steals a player contributed.
- PCT_TOV: % Percent of team's turnovers a player contributed.

Miscellaneous other statistics

- BLKA: Nnumber of field goal attempts by a player or team that was blocked by the opposing team. *(per minute)*
- OPP_PTS_2ND_CHANCE: Number of points an opposing team scores on a possession when the opposing team rebounds the ball on offense. *(per minute)*
- OPP_PTS_FB: Number of points scored by an opposing player or team while on a fast break. *(per minute)*
- OPP_PTS_OFF_TOV: Number of points scored by an opposing player or team following a turnover. *(per minute)*
- OPP_PTS_PAINT: Number of points scored by an opposing player or team in the paint.
- PFD: Number of fouls that a player or team has drawn on the other team. *(per minute)*
- PTF_FB: Number of points scored by a player or team while on a fast break. *(per minute)*

- PTS_2ND_CHANCE: Number points scored by a team on a possession that they rebound the ball on offense. *(per minute)*
- PTS_OFF_TOV: Number of points scored by a player or team following an opponent's turnover. *(per minute)*
- PTS_PAINT: Number of points scored by a player or team in the paint. *(per minute)*

BIBLIOGRAPHY

[1] Arianna Agosto, Giuseppe Cavaliere, Dennis Kristensen, and Anders Rahbek. "Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX)." In: *Journal of Empirical Finance* 38 (2016), pp. 640–663.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. "Optuna: A Next-generation Hyperparameter Optimization Framework." In: *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[3] Naomi S Altman. "An introduction to kernel and nearest-neighbor nonparametric regression." In: *The American Statistician* 46.3 (1992), pp. 175–185.

[4] Giovanni Angelini and Luca De Angelis. "PARX model for football match predictions." In: *Journal of Forecasting* 36.7 (2017), pp. 795–807.

[5] Giovanni Angelini and Luca De Angelis. "Efficiency of online football betting markets." In: *International Journal of Forecasting* 35.2 (2018), pp. 712–721.

[6] Kenneth Joseph Arrow. *Aspects of the theory of risk-bearing*. Yrjö Jahnssonin Säätiö, 1965.

[7] Gianluca Baio and Marta Blangiardo. "Bayesian hierarchical model for the prediction of football results." In: *Journal of Applied Statistics* 37.2 (2010), pp. 253–264.

[8] Rose D Baker and Ian G McHale. "Optimal betting under parameter uncertainty: Improving the Kelly criterion." In: *Decision Analysis* 10.3 (2013), pp. 189–199.

[9] Rose Baker and Philip Scarf. "Modifying Bradley–Terry and other ranking models to allow ties." In: *IMA Journal of Management Mathematics* 32.4 (2020), pp. 451–463.

[10] Jeremy Balka, Anthony Desmond, et al. "Kelly Investing with Iteratively Updated Estimates of the Probability of Success." PhD thesis. 2017.

[11] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[12] Daniel Berrar, Philippe Lopes, Jesse Davis, and Werner Dubitzky. "Guest editorial: special issue on machine learning for soccer." In: *Machine Learning* 108.1 (2019), pp. 1–7.

[13] Daniel Berrar, Philippe Lopes, and Werner Dubitzky. "Incorporating domain knowledge in machine learning for soccer outcome prediction." In: *Machine Learning* 108.1 (2019), pp. 97–126.

[14] Jose Blanchet, Lin Chen, and Xun Yu Zhou. "Distributionally robust mean-variance portfolio selection with Wasserstein distances." In: *Management Science* 68.9 (2022), pp. 6382–6410.

[15] Paul T Boggs and Jon W Tolle. "Sequential quadratic programming." In: *Acta numerica* 4 (1995), pp. 1–51.

[16] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. "Learning structured embeddings of knowledge bases." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 25. 1. 2011, pp. 301–306.

[17] Georgi Boshnakov, Tarak Kharrat, and Ian G McHale. "A bivariate Weibull count model for forecasting association football scores." In: *International Journal of Forecasting* 33.2 (2017), pp. 458–466.

[18] Ralph Allan Bradley and Milton E Terry. "Rank analysis of incomplete block designs: I. The method of paired comparisons." In: *Biometrika* 39.3/4 (1952), pp. 324–345.

[19] Leo Breiman et al. *Optimal gambling systems for favorable games*. 1961.

[20] Sid Browne and Ward Whitt. "Portfolio choice and the Bayesian Kelly criterion." In: *Advances in Applied Probability* 28.4 (1996), pp. 1145–1176.

[21] Enzo Busseti, Ernest K Ryu, and Stephen Boyd. "Risk-constrained Kelly gambling." In: *The Journal of Investing* 25.3 (2016), pp. 118–134.

[22] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. "A limited memory algorithm for bound constrained optimization." In: *SIAM Journal on Scientific Computing* 16.5 (1995), pp. 1190–1208.

[23] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 785–794.

[24] Dani Chu, Yifan Wu, and Tim B Swartz. "Modified Kelly criteria." In: *Journal of Quantitative Analysis in Sports* 14.1 (2018), pp. 1–11.

[25] Anthony C Constantinou. "Dolores: a model that predicts football match outcomes from all over the world." In: *Machine Learning* 108.1 (2019), pp. 49–75.

[26] Anthony C Constantinou and Norman E Fenton. "Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models." In: *Journal of Quantitative Analysis in Sports* 8.1 (2012), pp. 1559–0410.

[27] Anthony C Constantinou and Norman E Fenton. "Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries." In: *Journal of Quantitative Analysis in Sports* 9.1 (2013), pp. 37–50.

[28] Anthony C Constantinou, Norman E Fenton, and Martin Neil. “Profiting from an inefficient Association Football gambling market: Prediction, Risk and Uncertainty using Bayesian networks.” In: *Knowledge-Based Systems* 50 (2013), pp. 60–86.

[29] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[30] Martin Crowder, Mark Dixon, Anthony Ledford, and Mike Robinson. “Dynamic modelling and prediction of English Football League matches for betting.” In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 51.2 (2002), pp. 157–168.

[31] László Csató. “Coronavirus and sports leagues: obtaining a fair ranking when the season cannot resume.” In: *IMA Journal of Management Mathematics* 32.4 (2021), pp. 547–560.

[32] Pierre Dangauthier, Ralf Herbrich, Tom Minka, and Thore Graepel. “Trueskill through time: Revisiting the history of chess.” In: *Advances in neural information processing systems*. 2008, pp. 337–344.

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

[34] Werner Dubitzky, Philippe Lopes, Jesse Davis, and Daniel Berrar. “The Open International Soccer Database for machine learning.” In: *Machine Learning* 108.1 (2019), pp. 9–28.

[35] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[36] Edward S Epstein. “A scoring system for probability forecasts of ranked categories.” In: *Journal of Applied Meteorology* 8.6 (1969), pp. 985–987.

[37] Eugene F Fama. “Efficient capital markets: A review of theory and empirical work.” In: *The journal of Finance* 25.2 (1970), pp. 383–417.

[38] Eugene F Fama. “Market efficiency, long-term returns, and behavioral finance.” In: *Journal of financial economics* 49.3 (1998), pp. 283–306.

[39] Alistair D Fitt. “Markowitz portfolio theory for soccer spread betting.” In: *IMA Journal of Management Mathematics* 20.2 (2009), pp. 167–184.

[40] David Forrest, John Goddard, and Robert Simmons. “Odds-setters as forecasters: The case of English football.” In: *International Journal of Forecasting* 21.3 (2005), pp. 551–564.

[41] David Forrest and Robert Simmons. “Forecasting sport: the behaviour and performance of football tipsters.” In: *International Journal of Forecasting* 16.3 (2000), pp. 317–331.

[42] Egon Franck, Erwin Verbeek, and Stephan Nüesch. "Prediction accuracy of different market structures—bookmakers versus a betting exchange." In: *International Journal of Forecasting* 26.3 (2010), pp. 448–459.

[43] Jerome H Friedman. "Greedy function approximation: A gradient boosting machine." In: *Annals of Statistics* (2001), pp. 1189–1232.

[44] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[45] Mark E Glickman. "Parameter estimation in large dynamic paired comparison experiments." In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48.3 (1999), pp. 377–394.

[46] John Goddard. "Regression models for forecasting goals and match results in association football." In: *International Journal of forecasting* 21.2 (2005), pp. 331–340.

[47] Anjela Y Govan, Amy N Langville, and Carl D Meyer. "Offense-defense approach to ranking team sports." In: *Journal of Quantitative Analysis in Sports* 5.1 (2009).

[48] Anjela Y Govan, Carl D Meyer, and Russell Albright. "Generalizing Google's PageRank to rank national football league teams." In: *Proceedings of the SAS Global Forum*. Vol. 2008. 2008.

[49] Thore Graepel, Tom Minka, and R TrueSkill Herbrich. "A Bayesian skill rating system." In: *Advances in Neural Information Processing Systems* 19 (2007), pp. 569–576.

[50] Clive WJ Granger and M Hashem Pesaran. "Economic and statistical measures of forecast accuracy." In: *Journal of Forecasting* 19.7 (2000), pp. 537–560.

[51] Andreas Groll, Cristophe Ley, Gunther Schauberger, and Hans Van Eetvelde. "A hybrid random forest to predict soccer matches in international tournaments." In: *Journal of Quantitative Analysis in Sports* 15.4 (2019), pp. 271–287.

[52] Shengbo Guo, Scott Sanner, Thore Graepel, and Wray Buntine. "Score-based bayesian skill learning." In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2012, pp. 106–121.

[53] Maral Haghighat, Hamid Rastegari, and Nasim Nourafza. "A review of data mining techniques for result prediction in sports." In: *Advances in Computer Science: an International Journal* 2.5 (2013), pp. 7–12.

[54] Charles R Henderson. "Best linear unbiased estimation and prediction under a selection model." In: *Biometrics* (1975), pp. 423–447.

[55] Lisa Holton. "Is Markowitz Wrong? Market Turmoil Fuels Nontraditional Approaches to Managing Investment Risk." In: *Journal of Financial Planning* 22.1 (2009), p. 20.

[56] Kou-Yuan Huang and Wen-Lung Chang. "A neural network method for prediction of 2006 world cup football game." In: *The 2010 international joint conference on neural networks (IJCNN)*. IEEE. 2010, pp. 1–8.

[57] Lars Magnus Hvattum and Halvard Arntzen. "Using ELO ratings for match result prediction in association football." In: *International Journal of forecasting* 26.3 (2010), pp. 460–470.

[58] Z Ivanković, M Racković, B Markoski, D Radosav, and M Ivković. "Analysis of basketball games using neural networks." In: *Computational Intelligence and Informatics (CINTI), 2010 11th International Symposium on*. IEEE. 2010, pp. 251–256.

[59] Philippe Jorion. "Portfolio optimization in practice." In: *Financial analysts journal* 48.1 (1992), pp. 68–74.

[60] Raymond Kan and Guofu Zhou. "Optimal portfolio choice with parameter uncertainty." In: *Journal of Financial and Quantitative Analysis* 42.3 (2007), pp. 621–656.

[61] Dimitris Karlis and Ioannis Ntzoufras. "Analysis of sports data by using bivariate Poisson models." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 52.3 (2003), pp. 381–393.

[62] Dimitris Karlis and Ioannis Ntzoufras. "Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference." In: *IMA Journal of Management Mathematics* 20.2 (2008), pp. 133–145.

[63] John L Kelly Jr. "A New Interpretation of Information Rate." In: *The Bell System Technical Journal* (1956).

[64] James Kennedy and Russell Eberhart. "Particle swarm optimization (PSO)." In: *Proc. IEEE International Conference on Neural Networks, Perth, Australia*. 1995, pp. 1942–1948.

[65] Jon M Kleinberg. "Authoritative sources in a hyperlinked environment." In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[66] Siem J Koopman and Rutger Lit. "A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League." In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.1 (2015), pp. 167–186.

[67] Siem J Koopman and Rutger Lit. "Forecasting football match results in national league competitions using score-driven time series models." In: *International Journal of Forecasting* 35.2 (2019), pp. 797–809.

[68] Justin Kubatko, Dean Oliver, Kevin Pelton, and Dan T Rosenbaum. "A starting point for analyzing basketball statistics." In: *Journal of Quantitative Analysis in Sports* 3.3 (2007).

[69] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[70] Jiří Lahvička. "Using Monte Carlo simulation to calculate match importance: The case of English Premier League." In: *Journal of Sports Economics* 16.4 (2015), pp. 390–409.

[71] Henry Allen Latane. "Criteria for choice among risky ventures." In: *The Kelly capital growth investment criterion: theory and practice*. World Scientific, 2011, pp. 35–46.

[72] Verica Lazova and Lasko Basnarkov. "PageRank Approach to Ranking National Football Teams." In: *arXiv preprint arXiv:1503.01331* (2015).

[73] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series." In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[74] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[75] Gordon Leitch and J Ernest Tanner. "Economic forecast evaluation: profits versus the conventional error measures." In: *The American Economic Review* (1991), pp. 580–590.

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

[77] Christophe Ley, Tom Van de Wiele, and Hans Van Eetvelde. "Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches." In: *Statistical Modelling* 19.1 (2019), pp. 55–77.

[78] Bin Li and Steven CH Hoi. "Online portfolio selection: A survey." In: *ACM Computing Surveys (CSUR)* 46.3 (2014), pp. 1–36.

[79] Bernard Loeffelholz, Earl Bednar, and Kenneth W Bauer. "Predicting NBA games using neural networks." In: *Journal of Quantitative Analysis in Sports* 5.1 (2009).

[80] Leonard C MacLean, Edward O Thorp, Yonggan Zhao, and William T Ziemba. "Medium term simulations of the full Kelly and fractional Kelly investment strategies." In: *The Kelly Capital Growth Investment Criterion: Theory and Practice*. World Scientific, 2011, pp. 543–561.

[81] Leonard C MacLean, Edward O Thorp, and William T Ziemba. "Good and bad properties of the Kelly criterion." In: *Risk* 20.2 (2010), p. 1.

[82] Leonard C MacLean, Edward O Thorp, and William T Ziemba. *The Kelly capital growth investment criterion: Theory and practice*. Vol. 3. World Scientific, 2011.

[83] Leonard C MacLean, William T Ziemba, and George Blazenko. "Growth versus security in dynamic investment analysis." In: *Management Science* 38.11 (1992), pp. 1562–1585.

[84] Michael J Maher. “Modelling association football scores.” In: *Statistica Neerlandica* 36.3 (1982), pp. 109–118.

[85] Harry Markowitz. “Portfolio Selection.” In: *The Journal of Finance* 7.1 (1952), pp. 77–91.

[86] Philip Z Maymin. “Wage against the machine: A generalized deep-learning market test of dataset value.” In: *International Journal of Forecasting* 35.2 (2017), pp. 776–782.

[87] Peter McCullagh. “Regression models for ordinal data.” In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2 (1980), pp. 109–127.

[88] Ian McHale and Phil Scarf. “Modelling the dependence of goals scored by opposing teams in international soccer matches.” In: *Statistical Modelling* 11.3 (2011), pp. 219–236.

[89] Blake McShane, Moshe Adrian, Eric T Bradlow, and Peter S Fader. “Count models based on Weibull interarrival times.” In: *Journal of Business & Economic Statistics* 26.3 (2008), pp. 369–378.

[90] Richard O Michaud and Robert O Michaud. *Efficient asset management: a practical guide to stock portfolio optimization and asset allocation*. Oxford University Press, 2008.

[91] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality.” In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[92] Dragan Miljković, Ljubiša Gajić, Aleksandar Kovačević, and Zora Konjović. “The use of data mining for basketball matches outcomes prediction.” In: *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on*. IEEE. 2010, pp. 309–312.

[93] Tom Minka, Ryan Cleven, and Yordan Zaykov. “TrueSkill 2: An improved Bayesian skill rating system.” In: *Tech. Rep.* (2018).

[95] Edmund Noon, William J Knottenbelt, and Daniel Kuhn. “Kelly’s fractional staking updated for betting exchanges.” In: *IMA Journal of Management Mathematics* 24.3 (2013), pp. 283–299.

[96] Marco Ottaviani and Peter Norman Sørensen. “The favorite-longshot bias: An overview of the main explanations.” In: *Handbook of Sports and Lottery markets*. Elsevier, 2008, pp. 83–101.

[97] Alun Owen. “Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter.” In: *IMA Journal of Management Mathematics* 22.2 (2011), pp. 99–113.

[98] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab, 1999.

[99] Rodney J Paul and Andrew P Weinbach. "Does sportbook.com set pointspreads to maximize profits?" In: *The Journal of Prediction Markets* 1.3 (2007), pp. 209–218.

[100] Rodney J Paul and Andrew P Weinbach. "Price setting in the NBA gambling market: Tests of the Levitt model of sportsbook behavior." In: *International Journal of Sport Finance* 3.3 (2008), p. 137.

[101] Rodney J Paul and Andrew P Weinbach. "The determinants of betting volume for sports in North America: Evidence of sports betting as consumption in the NBA and NHL." In: *International Journal of Sport Finance* 5.2 (2010), p. 128.

[102] Andre F Perold. "Large-scale portfolio optimization." In: *Management science* 30.10 (1984), pp. 1143–1160.

[103] Ole Peters and Murray Gell-Mann. "Evaluating gambles using dynamics." In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26.2 (2016), p. 023103.

[104] Georg Ch Pflug, Alois Pichler, and David Wozabal. "The 1/N investment strategy is optimal under high model ambiguity." In: *Journal of Banking & Finance* 36.2 (2012), pp. 410–417.

[105] Keith Pilbeam. *Finance & financial markets*. Macmillan International Higher Education, 2018.

[106] Richard Pollard and Gregory Pollard. "Home advantage in soccer: A review of its existence and causes." In: *International Journal of Soccer and Science* 3.1 (2005), pp. 28–44.

[107] Keshav Puranmalka. "Modelling the NBA to make better predictions." MA thesis. Massachusetts Institute of Technology, 2013.

[108] Pieter Robberechts and Jesse Davis. "Forecasting the FIFA World Cup–Combining result-and goal-based team ability parameters." In: *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2018 workshop*. Vol. 2284. Springer. 2018, pp. 52–66.

[109] Brian M Rom and Kathleen W Ferguson. "Post-modern portfolio theory comes of age." In: *Journal of Investing* 3.3 (1994), pp. 11–17.

[110] Havard Rue and Oyvind Salvesen. "Prediction and retrospective analysis of soccer matches in a league." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 49.3 (2000), pp. 399–418.

[111] Paul A Samuelson. "The "fallacy" of maximizing the geometric mean in long sequences of investing or gambling." In: *Proceedings of the National Academy of sciences* 68.10 (1971), pp. 2493–2496.

[112] Paul A Samuelson. "Why we should not make mean log of wealth big though years to act are long." In: *The Kelly Capital Growth Investment Criterion: Theory and Practice*. World Scientific, 2011, pp. 491–493.

[113] Paul A Samuelson. "Proof that properly anticipated prices fluctuate randomly." In: *The world scientific handbook of futures markets*. World Scientific, 2016, pp. 25–38.

[114] William F Sharpe. "The sharpe ratio." In: *Journal of portfolio management* 21.1 (1994), pp. 49–58.

[115] Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A Smith. "Predicting the NFL using Twitter." In: *arXiv preprint arXiv:1310.6998* (2013).

[116] John G Skellam. "The frequency distribution of the difference between two Poisson variates belonging to different populations." In: *Journal of the Royal Statistical Society. Series A (General)* 109.Pt 3 (1946), pp. 296–296.

[117] ChiUng Song, Bryan L Boulier, and Herman O Stekler. "The comparative accuracy of judgmental and model forecasts of American football games." In: *International Journal of Forecasting* 23.3 (2007), pp. 405–413.

[118] Gustav Sourek, Vojtech Aschenbrenner, Filip Zelezny, Steven Schockaert, and Ondrej Kuzelka. "Lifted relational neural networks: Efficient learning of latent relational structures." In: *Journal of Artificial Intelligence Research* 62 (2018), pp. 69–100.

[119] Martin Spann and Bernd Skiera. "Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters." In: *Journal of Forecasting* 28.1 (2009), pp. 55–72.

[120] Herman O Stekler, David Sendor, and Richard Verlander. "Issues in sports forecasting." In: *International Journal of Forecasting* 26.3 (2010), pp. 606–621.

[121] Alec Stephenson and Jeff Sonas. *PlayerRatings: Dynamic Updating Methods for Player Ratings Estimation*. R package version 1.0-3. 2019.

[122] Erik Štrumbelj. "On determining probability forecasts from betting odds." In: *International Journal of Forecasting* 30.4 (2014), pp. 934–943.

[123] Qingyun Sun and Stephen Boyd. "Distributional Robust Kelly Gambling." In: *arXiv preprint arXiv:1812.10371* (2018).

[124] Edward O Thorp. "The Kelly criterion in blackjack sports betting, and the stock market." In: *Handbook of asset and liability management*. Elsevier, 2008, pp. 385–428.

[125] Alkeos Tsokos, Santhosh Narayanan, Ioannis Kosmidis, Gianluca Baio, Mihai Cucuringu, Gavin Whitaker, and Franz Király. "Modeling outcomes of soccer matches." In: *Machine Learning* 108.1 (2019), pp. 77–95.

[126] Matej Uhrín. "System for autonomous betting with optimal wealth allocation." MA thesis. Czech Technical University, Faculty of Electrical Engineering, 2018.

[127] Matej Uhrín, Gustav Šourek, Ondrej Hubáček, and Filip Železný. “Optimal sports betting strategies in practice: an experimental review.” In: *IMA Journal of Management Mathematics* 32.4 (2021), pp. 465–489.

[129] Jan Van Haaren and Guy Van den Broeck. “Relational learning for football-related predictions.” In: *Latest Advances in Inductive Logic Programming*. World Scientific, 2015, pp. 237–244.

[130] Vasiliĭ Grigorevich Voinov and Mikhail Stepanovich Nikulin. *Unbiased Estimators and Their Applications: Volume 1: Univariate Case*. Vol. 263. Springer Science & Business Media, 2012.

[131] Petar Vračar, Erik Štrumbelj, and Igor Kononenko. “Modeling basketball play-by-play data.” In: *Expert Systems with Applications* 44 (2016), pp. 58–66.

[132] Chris Whitrow. “Algorithms for optimal allocation of bets on many simultaneous events.” In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 56.5 (2007), pp. 607–623.

[133] Fabian Wunderlich and Daniel Memmert. “Are betting returns a useful measure of accuracy in (sports) forecasting?” In: *International Journal of Forecasting* 36.2 (2020), pp. 713–722.

[134] Albrecht Zimmermann, Sruthi Moorthy, and Zifan Shi. “Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned.” In: *arXiv preprint arXiv:1310.3607* (2013).

b

PUBLICATIONS OF THE AUTHOR

List of publications originally presented for the purpose of the dissertation defense. Citations from Web of Science, Scopus, and Google Scholar listed as of January 2024.

B.1 PUBLICATIONS RELATED TO THE TOPIC OF THE THESIS

B.1.1 *Journal papers*

Hubáček, Ondřej and Gustav Šourek. "Beating the market with a bad predictive model." In: *International Journal of Forecasting* 39.2 (2023), pp. 691–719
WoS: ∅, Scopus: ∅, Google: 6
Journal IF: 7.9

David Mlčoch and **Hubáček, Ondřej**. "Competing in daily fantasy sports using generative models." In: *International Transactions in Operational Research* 31.3 (2023), pp. 1515–1532
WoS: ∅, Scopus: ∅, Google: ∅
Journal IF: 3.1

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. "Forty years of score-based soccer match outcome prediction: an experimental review." In: *IMA Journal of Management Mathematics* 33.1 (2021), pp. 1–18
WoS: 3, Scopus: 4, Google: 9
Journal IF: 1.7

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. "Exploiting sports-betting market using machine learning." In: *International Journal of Forecasting* 35.2 (2019), pp. 783–796
WoS: 23, Scopus: 28, Google: 72
Journal IF: 7.9

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. "Learning to predict soccer results from relational data with gradient boosted trees." In: *Machine Learning* 108.1 (2019), pp. 29–47

WoS: 33, Scopus: 35, Google: 68
Journal IF: 7.5

Matej Uhrín et al. “Optimal sports betting strategies in practice: an experimental review.” In: *IMA Journal of Management Mathematics* 32.4 (2021), pp. 465–489
WoS: 1, Scopus: 2, Google: 10
Journal IF: 1.7

B.1.2 *Conference papers*

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. “Lifted Relational Team Embeddings for Predictive Sport Analytics.” In: *CEUR Workshop Proceedings*. Vol. 2206. CEUR-WS, 2018, pp. 84–91
WoS: ∅, Scopus: 1, Google: 2

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. “Score-based soccer match outcome modeling–an experimental review.” In: *Proceedings of MathSport International 2019 Conference*. 2019. URL: `http://www.mathsportinternational.com/MathSport2019Proceedings.pdf`
WoS: ∅, Scopus: ∅, Google: 9

B.2 OTHER PUBLICATIONS OF THE AUTHOR

B.2.1 *Conference papers*

Hubáček, Ondřej, Gustav Šourek, and Filip Železný. “Deep learning from spatial relations for soccer pass prediction.” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11330 LNAI. 2019, pp. 159–166
WoS: ∅, Scopus: 5, Google: 13