Machine Learning in Football Betting: Testing Profitability on the Betfair Exchange

Master′s Thesis 2025

Christian Beck Allerød & Kasper Vestergaard Bargisen

Aalborg University Business School
Economics

Economics
Aalborg University
http://www.aau.dk

Title:
Machine Learning in Football Betting: Testing Profitability on the Betfair Exchange

Participants:
Christian Beck Allerød
Kasper Vestergaard Bargisen

Project Group:
24

Supervisor:
Thibault Laurentjoye

Page Numbers: 63.3 standard pages

Date of Completion:
May 30^{th}, 2025

Abstract:

This thesis investigates whether machine learning can be used to construct profitable betting strategies on the Betfair betting exchange in the top 5 European football leagues. We extend previous work by Hubáček et al. (2019) by implementing an XGBoost model with a custom loss function designed to decorrelate predictions from market odds. Strategies are built using both the Kelly-Criterion and Modern Portfolio Theory (MPT), and evaluated with walk-forward cross-validation, and a validation and out-of-sample set to avoid overfitting.

Unfiltered strategies yield negative returns, but performance improves when applying simple rule-based filters; betting only when predicted probabilities exceed market-implied probabilities by 25%, and limiting odds to below 2.0. The best-performing strategy, based on MPT with the probability filter, achieves a 3.71% annual growth rate across our entire backtesting period.

Although 3.71% is a relatively modest yearly return, the strategy offers diversification benefits due to its low correlation with traditional assets and can be valuable for maximizing the payoff from bookmaker's freebets.

Contents

1 Introduction

For economists, markets are exciting. Behavioral economists and micro-economists focus on how market participants and market makers behave in the market. Econometricians and financial economists have an econometrical and statistical approach to understanding and forecasting the market. Game theorists want to identify the strategic interactions in the market and market economists focus on how a market functions.

Markets are a broad and multifaceted concept. Markets encompass a wide range of assets, goods and services. In the last decades, the development in technological modeling tools, computational power and ease of access has been exponential, leading econometric modeling and forecasting to be a prominent part of being an economist in the 21^{st} century. Oftentimes, research papers focus on markets concerning stocks, bonds, house prices, raw materials and energy, but markets also exist for less-institutionalized matters.

One such example is sports betting - a domain that exhibit the same fundamental characteristics as other markets but is mostly viewed as a mean of recreational entertainment. Here, market participants place wagers against bookmakers or other market participants on bet exchanges. Much like in traditional financial markets, the aggregate belief and expectation of the probability of a certain outcome to occur determines the price of a bet on bet exchanges. Specifically, bet exchanges offer a more clean and honest exchange of opposite positions in wagers for specific events, while bookmakers are free to set their prices as they wish. Betting markets offer a unique real-world testing ground for efficiency, human behavior, decision-making under uncertainty, risk-aversion among market participants and predictive models.

This thesis aims to investigate whether a machine learning-based forecasting model can generate positive returns in the long run when applied to betting exchange odds in European football. The project is therefore situated in the intersection of predictive modeling, investigating the efficiency of the football betting market and attempting to identify and exploit human behavior. The data is collected for the five major football leagues in Europe; the English Premier League, the Spanish La Liga, the German 1. Bundesliga, the Italian Serie A and the French Ligue 1 from 2017 to April 2025.

The project investigates the possibilities of generating positive returns in the 1X2 market[1] by building betting strategies by leveraging the predictions from the highly accurate and flexible XGBoost algorithm with a custom loss function. In order to build our betting strategies we implement common risk management techniques in the Kelly-Criterion and Modern Portfolio Theory.

Innate in the investigation of the possibilities of generating positive returns is an examination of the level of efficiency in the market. Fundamental metrics for the betting market changed during the Covid-19 pandemic, creating a change in regime between the pre-pandemic and the post-pandemic period, providing further opportunities to investigate changes in human behavior.

In order to assure robustness and generalizability of the results of the predictive model, the methodology include a rigorous process for avoiding backtest overfitting by employing both validation and out-of-sample tests as well as a strict walk-forward cross-validation procedure.

The thesis proceeds as follows: in chapter 2 we will give an introduction to odds and betting exchanges, as well as an initial exploration of efficiency in the sports betting market by especially exploring cognitive biases in the market. In chapter 3 we present the data, the risk management methods we are using to build our betting strategies and the measures to evaluate the results. Chapter 4 is more technical, as we go through how XGBoost works, which allows us to implement our own loss function. Additionally, we present which steps we have taken in order to avoid backtest overfitting. In chapter 5 we present, analyze and discuss our results, after which we conclude on our findings in chapter 6.

[1] The 1X2 market is a standard betting format in football, where: 1 represents the odds for a victory for the home team, X represents the odds for a draw and 2 represents the odds for a victory for the away team.

2 The Sports Betting Market

In the sports betting market, market participants, called *bettors*, place a stake on a specific pre-determined wager (*bet*). The bet is constructed based on the odds that the bet-providing party is willing to sell the wager at. Similarly to asset pricing in financial markets, the seller sells at a price that is perceived as acceptable. Therefore, one can view the odds of a wager as the price of the asset.

In sports betting, the odds is directly translated to the payoff per stake *if* the wagered outcome happens; the higher the odds (the lower the probability), the higher the payoff – and vice versa. If the wagered outcome does not occur, the bettor loses the whole stake. [1]

2.1 Odds and Implied Probability

There are three primary formats of betting odds; *decimal*, *fractional* and *American* odds. In this project, the **decimal** odds format will be used due to its prevalence and widespread adoption within European sports betting markets; Suppose a bettor enters a bet with a stake of 1.00€ and a potential payoff of 2.50€, the decimal odds are then 2.50, as the payoff is 2.50 times the size of the stake - implicitly manifesting a *profit* of 1.50€. (López, 2022, ch. 2) [2]

Odds are numerically based on the probability that an event occurs, denoted the

[1] Some bookmakers provide various mechanisms that refund losses, hand out freebets etc. Due to the possibility of changing such policies and keeping the framework as simple as possible, for now, we ignore these possibilities. However, we provide a brief discussion of how our strategy may be useful to exploit these policies in section 5.7.

[2] Equally, a bet with a stake of 1.00€ and a potential payoff of 2.50€, hence a potential profit of 1.50€; the **fractional odds** is: $\frac{3}{2}$ and the **American odds** is: +150. In both cases; instead of directly indicating the total *payoff* as the decimal odds do, it indicates the *profit*. (López, 2022)

implied probability, given by the following simple equation:

$$\text{Implied Probability} = \frac{1}{\text{decimal odds}} \tag{2.1}$$

, entertaining the above example; the decimal odds is 2.50, hence the implied probability is $\frac{1}{2.50} = 0.40$, i.e. the outcome with odds 2.50 is expected to happen four out of ten iterations. (López, 2022, ch. 4)

Note that the example above does *not* account for commission or transaction fees, which are applied by both traditional bookmakers and betting exchanges.

For bookmakers, the profit margin is embedded in the provided odds, while on betting exchanges, a commission is typically paid from profits of a wager. Bookmakers can freely price the provided bets, such that their advantageous position to the bettor is maintained through time. Therefore, the lower the profit margin, the higher the *payout rate*[3]. Some bookmakers even promote their platform by displaying their repayment rate - relative to the repayment rate of its competitors.

While bookmakers are free to provide the bets they favor, the provided odds among bookmakers must be very similar to the market odds. Otherwise, bettors would hypothetically be able to exploit the differences in odds between platforms and generate arbitrage by combining bets on opposite outcomes in the same events. Refer to section 2.4.2 for an investigation of such arbitrage opportunities.

In contrast, the odds on bet exchanges simply follow the market mechanisms of supply and demand. In the project, we use odds from betting exchanges rather than from bookmakers. For an overview of the structure and functionality of betting exchanges, refer to section 2.2.

In sports betting, bettors have the opportunity to wager on a wide range of possible outcomes - not just who wins the match. These outcomes can include the number of yellow cards, the total number of corners, the time of the first goal, which team scores first, which team is awarded more free-kicks, and many other specific events that may occur during a match.

Bettors can combine odds from different events in the same wager, denoted *parlays*. Suppose; two wagers with odds 2 are combined in a parlay, the odds are then $2 \times 2 = 4$. Oftentimes, bettors can even combine different non-contradictory

[3] The payout rate (repayment rate) is the guaranteed payout by betting on all possible outcomes (Hegarty & Whelan, 2024). Suppose for a single match, a home team victory has odds of 1.66, draw odds of 4.33 and away team victory odds of 4.50; the implied probability is then $\frac{1}{1.66} + \frac{1}{4.33} + \frac{1}{4.50} = 1.0556$, hence a payout rate of $\frac{1}{1.0556} \times 100 = 94,73\%$. Therefore, the payout rate quantifies the proportion of the stake that the bookmaker returns to the bettor in the long run.

wagers in the same event, e.g. in a specific match; i) the home team wins, ii) at least two yellow cards for the away team and iii) more than six corners in total.

In this thesis, we will not combine odds. Instead, we focus on betting on individual matches.

Importantly, odds are not static nor fixed; rather, the odds adjust dynamically both before and during the match in response to various factors. Pre-match odds change in response to team news, injuries, etc. In-play match odds change rapidly in response to real-time events, such as goals, penalties, red cards, injuries and match momentum. In addition to this, market movements influence the odds, similarly to the continuous change in prices of stocks when demand and supply are not in equilibrium. Therefore, the odds on betting exchanges change continuously, and consequently, the bookmaker's odds will change. (Cortis & Levesley, 2016, ch. 2 & 6)

However, Cortis & Levesley (2016) have found that, while the bookmaker's odds broadly follow the market odds, they manipulate their provided odds in order to exploit inefficient situations in the demand for odds. This is a tool for maximizing profits - or at least ensuring long-term profitability.

2.2 Betting exchanges: Structure and Functionality

Traditional bookmakers are the direct counterpart to bettors as they act as the opposing party in a wager. Bookmakers provide odds on their own terms of risk management strategies, and the bettor is free to participate in the wager, however *never* as the role of the bet provider.

Bet exchanges operate as any peer-to-peer marketplace, i.e. a marketplace between equal bettors. Bettors do not participate in a wager *against the house*, instead, the wager is engaged between bettors taking opposite positions of the outcome of a specific event. On betting exchanges, it is *solely* the forces of supply and demand that dictate the prices of odds. No market participant acting as the bet provider in the betting market have the freedom to price the odds according to the ensurance of long-term profitability. Similarly, trades in the stock market are engaged between investors that have opposing expectations of outcomes, and the prices is determined by the aggregate demand and supply of the specific stock.

Due to these structural differences, betting exchanges are said to offer more efficient pricing. (Franck et al., 2010)

The market participant taking the role of the traditional bookmaker places a *lay bet* and the bettor placing a wager on an event occurring places a *back bet*. Like the stock market, betting exchanges operate on an order-matching system, that ensures that both lay bets and back bets are engaged at a jointly agreed odds. If the market odds are not priced correctly according to a market participant, the market participant can supply a lay bet or buy a back bet as a limit order[4]. Oftentimes betting exchanges charge a commission on *profits* from a wager, rather than having incorporated an embedded profit margin in the odds.

While Franck et al. (2010) argue that the prices on betting exchanges *incorporate the relevant news swiftly and fully, indicating a high level of efficiency*, this present project will be investigating the possibilities of creating long-term profits on the basis of bet exchange odds data. The innate institutional profit-maximizing strategic approach and the immense data-handling capabilities of the bookmakers is assumed to be a *greater* hurdle, than the efficiency of the bet exchange odds. Additionally, if we were to successfully generate systematic profits from a bookmaker, they would be free to exclude us at any time. Moreover, the project aims at investigating the level of efficiency of the sports betting market, which is difficult when utilizing the provided odds from bookmakers. In section 2.4, the level of efficiency in the betting market is discussed - and in section 3.1, the used data is presented.

2.3 Market Participants and Liquidity

There are notable differences in motivation behind participation in the financial markets and in the sports betting market. While recreational stock purchasing is observed, its prevalence remains relatively limited in the aggregated stock market. Conversely, such recreational purposes are much more prevalent among market participants in the sports betting market. This suggests that the appeal of sports betting is, relative to various other financial assets, more a form of entertainment or leisure.

Justifiably, one can imagine, this tendency can be attributed to the fundamental structure of the immediacy of outcomes, as well as the rich historical passion and even tribalism especially in the context of *European* football. In the paper Bruce

[4] A limit order; the placement of a bet at another price (odds) the market price (odds). Suppose that the odds for an outcome is 2.00, but the market participant believes the odds should be 1.90 – the market participant can then place a back bet at odds 1.90, and by default; the wager is only engaged *if* the odds are 1.90

et al. (2012), bettors are divided into two main groups of bettors: those who bet for recreational purposes (*recreational bettors*) and those whose bet based on a systematic and pre-defined strategy aimed at long-term profitability (*professional bettors*).

Bruce et al. (2012) argue that the presence of professional bettors, who generally hold more liquidity, can lead to sharper and more efficient odds, as they continuously exploit slight inefficiencies. These bets of substantial stake size are wagered at *large* sports events (such that individual bets do not move the odds) and the bets are based on strategies of optimizing mean return and aggregated returns. In contrast, the recreational bettor tends to view betting as a secondary activity and is more likely to place small-stake bets based on irrational grounds.

2.4 Efficiency in the Sports Betting Market

In the literature on betting market efficiency, the market is generally found to operate efficiently in line with the *Efficient Market Hypothesis* (EMH). While inefficiencies may arise in isolated cases, market mechanisms facilitate adjustments, ensuring that the market quickly corrects itself in accordance with EMH.

The betting market shares several key similarities with the financial markets, particularly in terms of efficiency, participant behavior, and risk management principles. In both markets, prices reflect publicly available information. On financial markets, prices adjust based on shifts in the investor's behavior and expectations in response to new data, financial reports, or policies. Similarly, prices in the betting market fluctuate according to team performance, injuries, and the expectations of the bettors.

2.4.1 Efficient Market Hypothesis

In the paper Fama (1970), the Efficient Market Hypothesis (EMH) is presented, and a comprehensive review thereof is performed. EMH stands as a cornerstone of financial thought and asset pricing till this day, arguing that in an efficient market, the prices always fully reflect the available information, and therefore; it is impossible to systemically make excessive profits in the market by trading on the basis of information readily available to the market.

Therefore, when new information is available in the market, the *new* information is immediately incorporated into the asset prices. Later, economist Burton Malkiel described the evolution of stock prices as a Random Walk, in his book *A Random Walk Down Wall Street* (Malkiel, 1973), referring to the unpredictable nature of the next new information – and therefore; the innate impossibility of consistently *"beating the market"*.

Fama identifies *three* forms of efficiency in the market: *weak*, *semi-strong* and *strong*. The three definitions of the degree of efficiency differ in the amount of information assumed to be reflected in asset prices.

At the **weak** form efficiency, the asset price reflects all *past* market data, which includes historical prices. At **semi-strong** form efficiency, the asset price reflects all publicly available information, including past market data, as well as fundamental contemporaneous information in public, such as company earnings and stock splits. Finally, the **strong** form efficiency, where asset prices reflect all the above *and* private information. Implicitly, within this form of efficiency, even insider-trading is futile, as private information is already embedded in the asset price. (Fama, 1970)

In the context of the Sports Betting Market, the same logic can be applied; at the **weak** form efficiency, the price of the bet (odds) will reflect all historical odds. At the **semi-strong** form efficiency, the odds will reflect all publicly available information, in the context of football; announced injuries, winning/losing streaks, importance of the match etc. At the **strong** form efficiency, the odds reflect all public *and* private information; hidden injuries, the manager's match-specific tactical philosophy, internal conflicts etc.

Intuitively, the strong form efficiency is somewhat unthinkable; in the financial markets, it is illegal for an employee to trade stock based on non-public information. Similarly, a football player is not allowed to bet on their own matches. In the real-world, this strong degree of efficiency is largely theoretical. Otherwise, in the perfect case of strong form efficiency in the market, the announcement of a financial report from a company would not have an impact on the stock price - and similarly, the announcement of injuries for the three most vital players of the team would not change the provided odds.

In Fama (1970), the actual degree of efficiency is investigated by continuously providing more information to a trading strategy and seeing *when* the information indeed would allow the trading strategy to generate profits consistently.

Fama found no significant evidence contradicting the semi-strong form efficiency, and hence neither the weak form efficiency, while the results were not as clear in the context of strong form efficiency. As Malkiel proposed; there is a lot of support in that changes in prices follow a Random Walk Model. The data for common stock, which Fama investigate, experience (very) close-to-zero positive autocorrelation, but it is argued that potential profits are absorbed by even minimum transactions costs for the investor attempting to exploit the small inefficiency. These transaction costs are not present in the proposed preferred strategy; buy-and-hold.

Although the definitions of efficiency in regards to the reflected information is applicable to the sports betting market, the fundamental nature of betting on sport matches is different to e.g. trading stocks. In the context of sports betting, there is no long-term buy-and-hold strategy - the act of *"investing"* in a game of sports has a predetermined time-frame, as is the case with options. Similarly, the bettor is able to close the deal prematurely, or hold the bet until expiration. Each bet is a discrete event with a defined outcome and time-frame - and if you hold the bet until expiration, you can either win a predetermined amount or lose the entire stake.

2.4.2 Inefficiencies and Market Imperfections

In the paper Deschamps (2007), the efficiency of the European football betting markets is examined, and possible arbitrage opportunities are identified. In an inefficient market, the participants will be able to systematically exploit the discrepancies in odds and obtaining risk-free profit. This is investigated by separately modeling i.) the best odds for each bet across up to 79 platforms and ii.) the average odds for each bet up to 79 platforms.

Across 6315 football mtaches in England, France, Germany, Italy, Spain and Scotland for the 2005/06 season, even the best possible odds across all platforms generate a negative return. Although Deschamps found that 6% of the games in the sample have potential arbitrage opportunities based on the best odds across platforms, inefficiencies are quickly corrected as market participants exploit them. This limits the viability of long-term arbitrage strategies for the strategic bettor in the aggregated European football betting market. Generally, the bookmaker prices the bets at market value – outlier odds are rare[5].

Deschamps investigates the presence of arbitrage opportunities in a standardized way, that aligns with the theory of Risk Management Buchdahl (2003, ch. 4); if the

[5] Outlier odds are odds that deviate significantly from the market average. The likelihood of experiencing arbitrage opportunities increases as the number of outlier odds increase. Suppose that a platform provides outlier odds for the home team to win in a specific match, while another platform provides outlier odds for the away team to win.

guaranteed payout exceeds the total stake, there is an arbitrage opportunity, i.e. risk-free profit. The *overrounding equation* is given by:

$$\lambda(i) = \sum_j \frac{1}{o_{ij}} - 1 \tag{2.2}$$

, where o_{ij} is the odds of outcome j, hence $\lambda(i)$ is the sum of the reciprocal of the odds of match i minus one, denoted the *overrounding of match i*. In a game of football, there are three possible outcomes in terms of the result; the home team wins, the away team wins or they draw, so ($j = 1, 2, 3$). In betting terms, respectively; 1, 2 and X.

If the sum of the reciprocal of the odds minus 1 for the three outcomes is equal to zero, it is denoted as *break-even*. If $\lambda(i) > 0$, there is no arbitrage, while if $\lambda(i) < 0$, there is an arbitrage opportunity, as the guaranteed payout exceeds the stake.

On February 15th 2025, for an English Premier League football match between Leicester and Arsenal, the provided 1X2 (1: home team win, X: draw and 2: away team win) odds from bet365 are 9.50, 5.25 and 1.33 in favor of Arsenal. According to equation 2.2, these odds generate an overrounding of 0.04762[6], and therefore no arbitrage. Suppose that the odds for an Arsenal victory increases to 1.43, all else equal, the overrounding would then be -0.00496[7], revealing an arbitrage opportunity.

Suppose that a bettor places a total of $100DKK$ on the match; $10.58DKK$ on outcome 1 at odds 9.50, $19.14DKK$ on outcome X at odds 5.25 and $70.28DKK$ on outcome 2 at odds 1.43[8]. No matter the outcome, the payoff is greater than the total stake of $100DKK$[9]. Note, in the above illustration, there are no transaction costs.

The arbitrage test presented in Deschamps (2007) is closely related to the nature of exploiting arbitrage opportunities in financial markets, presented in Hillier et al. (2012); in some inefficient situations, in the real-world, it is possible to generate guaranteed profits in the market by combining different negatively correlated assets or by simultaneously trading assets in different markets. Exploiting arbitrage

[6] $\lambda(LA_1) = \frac{1}{9.50} + \frac{1}{5.25} + \frac{1}{1.33} - 1 = 0.04762$

[7] $\lambda(LA_2) = \frac{1}{9.50} + \frac{1}{5.25} + \frac{1}{1.43} - 1 = -0.00496$

[8] Odds 1: $\frac{\frac{1}{9.50}}{\lambda(LA_2)+1} = 10.58DKK$, Odds X: $\frac{\frac{1}{5.25}}{\lambda(LA_2)+1} = 19.14DKK$, Odds 2: $\frac{\frac{1}{1.43}}{\lambda(LA_2)+1} = 70.28DKK$

[9] Odds 1: $10.58DKK \times 9.50 = 100.51DKK$
Odds X: $19.14DKK \times 5.25 = 100.49DKK$
Odds 2: $70.28DKK \times 1.43 = 100.50DKK$

in this pure form is the utmost *safe* trade one can make - but equally, it is a rare occurrence.

2.4.3 Favorite strategy vs. Long-shot strategy

In addition to Deschamps (2007), the conclusion that arbitrage opportunities appear but are quickly exploited and *closed* is concluded in Winkelmann et al. (2024). Furthermore, Deschamps (2007) found that the long-term yield is substantially higher when consistently betting on the favorite team due to *long-shot bias*. This was also investigated and concluded in Direr (2012), who even found that the profits are 4.45% and 2.78% from 2000 to 2011 when combining odds from the best odds across numerous bookmakers and when averaging odds across bookmakers, respectively. This result is based on *only* betting on teams that have a win probability of at least 90%[10].

The term *long-shot bias* from behavioral economics is a cognitive human tendency to overvalue and overestimate the probability of a highly improbable event of happening. This bias is very clear and applicable in the context of sports betting, and manifests itself as a disproportionate amount of money being wagered on outcomes with exceptionally high odds of happening - overruling a rational assessment of probability. Therefore, for such bets, the payoff is substantial, and the irrationality especially applies to speculative or recreational bettors.

The bias results from the inability of the human mind to comprehend the actual probability of an unlikely outcome. In the paper Kahneman & Tversky (1979), it is compared to the human ability to evaluate changes or differences; it is hot in the summer, but how hot is it? The music is loud, but how loud is it? Equally; the probability is low, but how low is it? Instead, the bettor sees the potential for an abnormal payoff. Kahneman & Tversky argue that the relative preference of betting on long-shots is *"...leading highly probable gambles to be under-priced"* due to market conditions. These conclusions resonate with the findings of both Direr (2012) as well as Snowberg & Wolfers (2010) - the latter did *not* generate profits by betting on the probable outcomes but proves the rate of return is substantially greater than when betting on the long-shots[11].

[10] The conclusion is based on investigating data that cover "21 championships" in eleven European countries from 2000 to 2011, consisting of 79.446 football matches. The odds are derived from six to ten bookmakers, summing to approx. 1.8 million odds. (Direr, 2012)

[11] The paper investigates the mean rate of return from consistently betting based on a simple strategy: Bet on all outcomes with odds $> x, x \in \mathbb{R}$. The analysis is based on odds in American horse racing from 1992 to 2001. The data consists of 6.4 million horse race starts, and they find that

Ambiguously, Franck et al. (2010) points out that evidence supports that the tendency to overvalue underdogs (underprice favorites) is less detectable on bet exchanges relative to the odds provided by the bookmakers, indicating greater efficiency on bet exchanges. Therefore, if the objective of this project was to solely exploit favorite bias, we would be better off betting on bookmaker platforms, and oppositely, if the objective was to solely exploit long-shot bias, we would be better off betting on a bet exchange.

According to Cortis & Levesley (2016), bookmakers are manipulative of the provided odds. They tend to provide *"too low"* odds for underdogs (hence; *too high* odds for favorites), according to the elasticity of the bettors.[12] It is a calculated endeavor to exploit the widespread human inability to evaluate low probabilities. At the end of the day; whether the decimal odds for a defender to score an own goal in the 5th minute with his left foot outside the penalty area is 350 or 700, the recreational bettor will enter the wager either way.

To test the above long-shot bias hypothesis, we look at bet exchange odds and available fixture data from the English Premier League, French Ligue 1, German 1. Bundesliga, Spanish La Liga and Italian Serie A between the 2016-17 season until the end of the 2023-24 season. We exclude the testing data (starting from the beginning of the 2024-25 season) in the below experiments, in order to prevent any data-snooping (see section 4.3.2).

We evaluate the betting strategies based on the *Cumulative PnL*, which is given by:

$$\text{Cumulative } PnL_i = \sum_{i=1}^{k} PnL_i \tag{2.3}$$

and

$$\begin{aligned} &\text{Wager won:} && PnL_i = (o_i - 1) \times b_i \times c \\ &\text{Wager lost:} && PnL_i = -b_i \end{aligned} \tag{2.4}$$

, where b_i and o_i is the stake and odds for the specific event i, respectively. $1 - c$ is the commission, which at Betfair is 5% of all profits, and therefore $c = 0.95$. PnL (Proft-and-Lossses) can therefore be both positive and negative, but unlike traditional financial markets, the PnL is never zero, as we either win or lose the bet.

the larger odds, the lower the mean rate of return in the long term. (Snowberg & Wolfers, 2010)

[12] The non-strategic bettors looking for a big payoff, which we have called recreational bettors, have an inelastic demand curve, while for strategic bettors, the curve is elastic. Therefore, while recreational bettors are insensitive to changes in odds, the strategic bettors are very sensitive to changes in the provided odds. The bookmakers are forced to have this in mind when pricing each bet. (Cortis & Levesley, 2016, ch. 6.2.2)

In figure 2.1, we test the long-shot bias by consistently placing a unit bet $b_i = 1$ on the favorite team with an odds of winning of 1.09 or less, which generates a profit of 2.36 over the period. However, across approx. 14,300 matches, only 48 instances fulfill this simple, yet strict requirement, and therefore there were only made 48 bets, one of which was lost.

If the acceptable interval of odds is marginally increased to 1.10, the profits are negative at -0.547. The outcomes of these two *favorite* strategies are illustrated in figure 2.1. Similarly to applying the benchmark odds of 1.09, the benchmark of 1.10 consists of *only* 75 total bets - six of which are losses that eliminate all profits. Further, if the benchmark odds are increased to 1.12, the losses are now -3.189. Evidently, in this simple strategy, all lost bets are very expensive, as one loss will set the bettor back significantly, potentially even eliminating *all* profits.

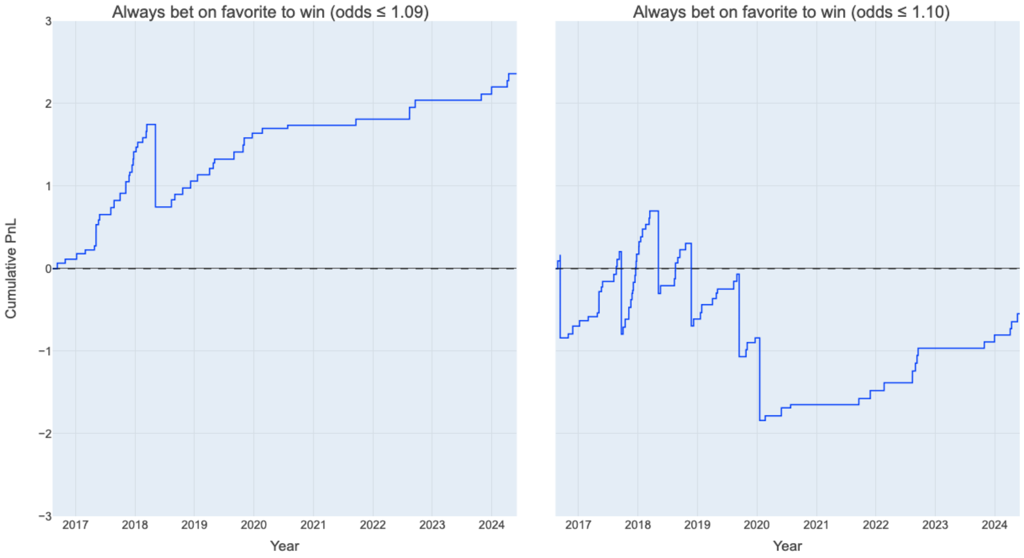


Figure 2.1: Favorite strategy: The cumulative PnL (Profit & Loss) of two simple strategies; always bet on a team with a winning odds of 1.09 and 1.10 or less, respectively: The figures illustrate the cumulative profits and losses (PnL) from the beginning of the 2016-17 season until the end of the 2023-24 season by implementing two very simple and similar strategies, but with very different results. On the left; in all possible matches, place a unit bet on teams with winning odds of 1.09 or less, and on the right; in all possible matches, place a unit bet on teams with winning odds of 1.10 or less. The figures are constructed based on all matches in the best league in England, France, Spain, Germany and Italy from the 2017-18 season until the end of the 2023-24 season. The odds are the Betfair bet exchange odds *before* the match starts.

Contrary to the evidence of the favorite strategy, Dixon & Pope (2004) find evidence of favorite *bias* (i.e. *reverse* long-shot bias) in English football matches between 1993 and 1996. They find that low-probability outcomes are priced too generously, and high-probability outcomes are priced too conservatively. This contrasts with the above-mentioned view that long-shot bias drives the price for low-probability outcomes down (i.e. drives the provided odds up).

We find that when betting a unit stake on teams with a bet exchange pre-game odds of at least 12.00, the period generates a profit of 22.73. However, the drawdowns are substantial; the maximum profits in the period is 129.25 and the minimum is a loss of -187.73, see figure 2.2. Clearly, when examining the graph on the left, one can argue that it is no more than a coincidence that the cumulative result is positive at the end of the examined subset of the data. Since January 2023, the PnL has experienced a downward trend - would the PnL be negative by January 2025? Who knows. Suppose that the determined benchmark odds is changed to 9.50, the cumulated result is instead -91.76 with a maximum of 54.89 and a minimum of -255.47, see 2.2 on the right. The number of bets in the two long-shot strategies are 1267 and 1956, respectively.

The difference between the *favorite team strategies* (figure 2.1) and the *long-shot strategies* (figure 2.2) is substantial.

Firstly, the level of volatility of the PnL is significantly higher in the long-shot strategies, compared to the favorite team strategies. The high volatility indicates the high risk the bettor faces, as its results are highly unpredictable and sensitive to the PnL. In figure 2.2, the strategy provides extreme profits in the latter part of the calendar-year 2019. Oppositely, the favorite team strategies demonstrate gradual profits, but great costs associated with losing a single bet.

Secondly, the number of betting opportunities. The lack of opportunities, especially for the favorite strategies, fundamentally makes it tough to generate substantial profits. The long-shot strategies, on the other hand, has a significantly higher frequency but also carry much greater uncertainty, where only a handful of wins can turn a large negative result into a situation with profit.

While the two simple strategies generate profits at certain thresholds, it must be underlined that the results are very uncertain. Providing thresholds and models for profits ex-post are generally valueless. How can a strategic bettor not be concerned if the downward trend in the long-shot strategy continues? Does a strategic bettor fancy that their favorite strategy is so dependent on outcomes of individual matches?

Nonetheless, it is clear that by implementing simple strategies, it is indeed pos-

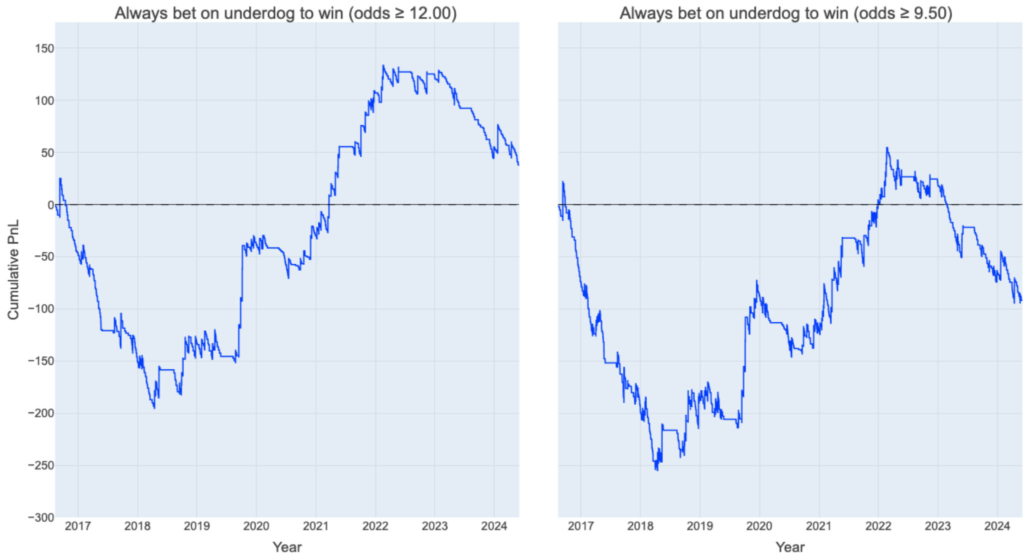


Figure 2.2: Long-shot strategy: The cumulative PnL (Profit & Loss) of two simple strategies; always bet on a team with a winning odds of at least 12.00 and 9.50, respectively: The figures illustrate the cumulative profits and losses (PnL) from the beginning of the 2016-17 season until the end of the 2023-24 season by implementing two very simple and similar strategies, but with very different results. On the left; in all possible matches, bet a unit stake on teams with winning odds of at least 12.00, and on the right; in all possible matches, bet a unit stake on teams with winning odds of at least 9.50. The figures are constructed based on all matches in the best League in England, France, Spain, Germany and Italy from the 2017-18 season until the end of the 2023-24 season. The odds are the BetFair bet exchange odds *before* the match starts

sible to generate profits on the betting market, i.e. the market is not perfectly efficient. Biases and inefficiencies persist, and therefore, in theory, it is *possible* to exploit such biases and inefficiencies and generate profit.

Based on the findings in figure 2.1 and 2.2, it is clear that different researchers find different results in different time-periods, supporting both a favorite and a longshot strategy. To us, out of the above, the favorite strategy with 1.09 as benchmark odds has provided the safest betting strategy for the present experiment. The strategy shows a consistent and stable trend in cumulative PnL, but it is always only *a few* lost bet away from being an aggregated loss. Additionally, in the figure, it seems that the market has acted on this inefficiency as bets with odds 1.09 or less have become significantly less frequent since Covid-19 in 2020. Refer to section 5.2 for a discussion thereof.

Ideally, we aim for a less restrictive strategy, allowing for more bets to be included, which in turn would reduce the dependency and reliance on the outcomes of individual bets. In essence; the utopian objective is to base a strategy on conditions that generates a cumulative PnL curve with a linear-like upward-sloping trend and a high frequency of bets.

2.4.4 Cognitive Biases and Behavior

In addition to the long-shot bias presented by the *Prospect Theory* from Kahneman & Tversky, various other irrational tendencies remain in the psychology of the human species, ever since the hunter-gatherer community thousands of years ago. All inconsistencies in objective rationality are interesting to the strategic bettor today. Sports betting is not solely about statistical probabilities: far-reaching irrational tendencies among bettors will create irrational betting patterns, which move the market, and possible inefficient prices might be exploitable.

The sports betting market operates at the intersection of statistical probabilities, regular financial market mechanisms and risk-management as well as irrational human behavior.

Overconfidence bias

In *Sapiens: A Brief History of Humankind*, Harari (2011, ch. 4), the author explains the human dependency on ancient myths and beliefs, making their perception of ability to hunt game or of the tribe to conquer new lands irrationally overestimated. Millennia later, we find market participants overestimating their own ability to predict the future of the markets and, perhaps in our case, the winning probability of the local semi-professional football team in the derby against the full-time professional team from the neighboring town. This psychological tendency among humans is denoted *overconfidence bias*.

Some market participants find themselves overly well-equipped in *beating the market*, which leads to excessive risk-taking: A *Self-valuation* bias that stems from past successes or selective memory, putting too much emphasis on past wins relative to past losses. Rationally, actual probability of outcomes cannot justify such confidence. (Wilkinson & Klaes, 2017, ch. 4)

The increased risk-taking in the context of betting might lead to increased betting volume, and the perception of a “safe bet” might not be based in reality.

Loss Aversion

In addition to identifying the Long-shot bias, Prospect Theory by Kahneman & Tversky also identified the loss aversion among humans. Humans tend to feel the pain of loosing more than the joy of equivalent wins, i.e. the disutility of being the loser in a given event exceeds the utility of being the winner in the same event. (Wilkinson & Klaes, 2017, ch. 3)

However, Shang et al. (2021) found that the loss aversion is greater when the market participant views the bet as an investment than as a recreational activity.

The heavy disutility of losing among humans, combined with the *laissez-faire* state of mind for recreational bettors might lead market participants to inadvertently chase losses. Clearly, losses are best equalized by winning long-shots – something that occurs rarely.

Gambler's Fallacy & Hot-Hand Fallacy

As presented in (Wilkinson & Klaes, 2017, ch 4), the gambler's fallacy is the irrational belief that past events influence future outcomes. A popular example of this is with roulette; if the ball has landed on red five times in a row, how can it possibly land on red again? Equally, If a football team has lost five matches in a row, are they not "due" for a win? These thoughts are based on a belief in the existence of non-existing patterns.

Contrarily to the Gambler's Fallacy is the Hot-Hand Fallacy. Bettors believe that the winning streak of a team is more likely to continue than it is to end. If a football team has won five matches in a row, how can they not win the next?

Both fallacies can lead to misinformed wagers, determined by passion and subjective belief – rather than on probability. Bettors have a tendency to view the above-mentioned football teams as "due for a win" and "unstoppable". It would be reasonable to test if there is profitability in constructing a strategy opposing these fallacies. We test the Gambler's Fallacy and the Hot-Hand Fallacy in section 2.4.5.

Emotional betting

Among recreational bettors, it follows naturally to conclude that personal preference and emotions influence the betting behavior. E.g. such bettor might *always* bet on their favorite team to win, or always bet on the rivals of their favorite team to lose – even though statistical analyses suggest otherwise.

Suppose this is a tendency among supporters of a big club with a large fanbase; this tendency would lower the odds for the team to win, and therefore increase the odds for the *other* team to win. Perhaps, the odds for a victory for monstrous clubs such as Manchester United and Real Madrid are *always* slightly lower than they *should* be according to the efficient market. Is it profitable to construct a betting strategy based on betting against the biggest clubs? We test this in section 2.4.5

2.4.5 Exploiting Cognitive Biases and Behavior

In addition to investigating the possibilities of exploiting inefficiencies in the market odds for favorite and long-shot strategies, which is a widely covered topic in various literature, we attempt to exploit systematic inefficiencies based on cognitive biases, emotions and match result momentum.

First, we investigate the possible inefficiencies that are systematic over time. In 2.4.3, we have already tested a possible systematic inefficiency by setting up a threshold for both favorite strategies and long-shot strategies. The pure mathematical conclusion is that we indeed did generate profits with a benchmark of less than or equal to 1.09 and greater than or equal to 12.00 in the English Premier League, but we also discovered that result would be negative if increased and decreased to the arbitrary number of 1.10 and 9.50, respectively.

Suppose we set up an equally simple threshold that allows for more observations (at least more than 48 observations in approx. 8 years, as the above-mentioned favorite strategy had), and the strategy generates a somewhat steadily increasing cumulative PnL curve. This would increasingly eliminate the risk of losing the entire winnings due to a single lost bet, and statistically, the increased number of observations would speak in favor of the actual legitimacy of the strategy in the long term.

As mentioned in section 2.4.4, suppose that bets tend to be overpriced for either home or away matches for specific teams. We find that by continuously betting on Arsenal to win all home matches generates a profit of 16.28 by continuously betting a stake of 1.

Historically, Arsenal is a great English club, but so is Liverpool and Manchester City, and performing the same investigation for these clubs generate *smaller* profit for the Liverpool-strategy of 4.18 and a loss for the Manchester City-strategy of -2.99. The cumulative PnL of the three clubs are depicted in figure 2.3. All clubs have not been relegated to the second-best division in England, and despite the results, both Liverpool and Manchester City have won more home matches between the 2016-17 and 2023-24 seasons than Arsenal.[13] Therefore, for the equation to make sense, the provided odds must be lower for Liverpool and Manchester City: The average odds for a home match victory for Arsenal is 1.95, 1.53 for Liverpool and 1.31 for Manchester City in the same period. However, from this simple investigation, it seems that the implied probability for a home victory for Arsenal is continuously underestimated.

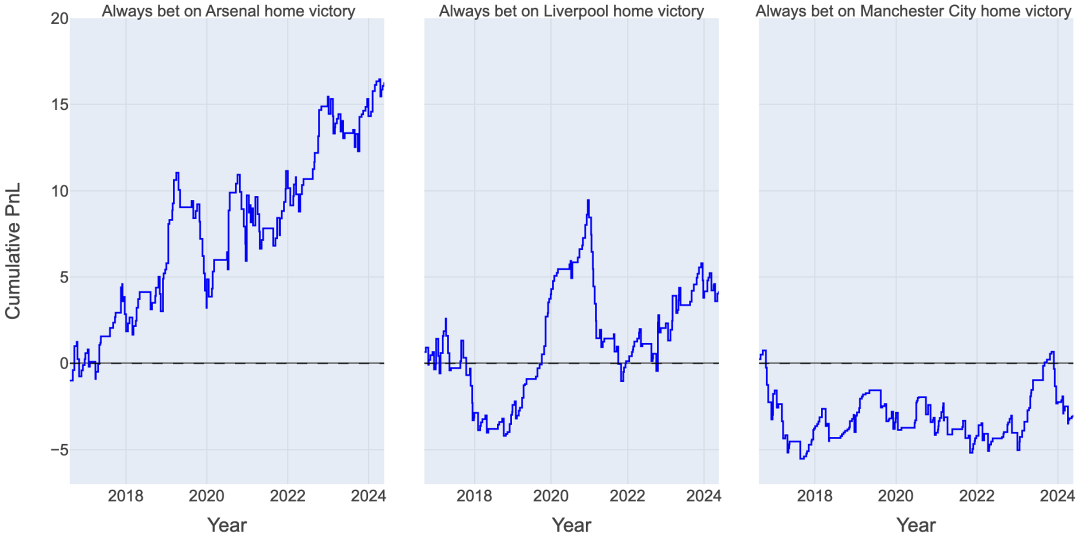


Figure 2.3: Always bet on a victory for the home team for specific clubs: The figures illustrate the cumulative profits and losses (PnL) from 2016 to 2024 for a simple strategy: Always bet on the home team - the home team being Arsenal, Manchester City and Liverpool, respectively. The figures are constructed based on all Premier League matches from the 2016-17 season until the end of the 2023-24 season. The odds are the average supplied odds by the bet supplier *before* the match, and each bet has a stake of 1.

Implementing this simple strategy on Liverpool indeed generates profits, but until 2020, the returns are negative as it skyrockets to a profit of almost 10 before plummeting back into the negative. Variance is high, and one could argue; if the data started a year later, the results would be significantly better and if we made this investigation in the end of 2020, the results would be even better. The variance is smaller for the implementation of the strategy for Manchester City, and it seems there is a clear and consistent trend.

For the implementation of the strategy for Arsenal, the trend is positive, and

[13] Between the 2016-17 season and the 2023-24 season, Arsenal have won 102 home matches, Liverpool 110 matches and Manchester United 117 wins - out of a total of 152 home matches.

it is clear that lost bets does not change the trend or result fundamentally, as we see for the strategy for Liverpool. To some extent, the curve can be explained as a positively-sloped linear curve with 152 observations, see figure 2.3.

Under the restrictions of logic and uncertainty about the future, the Arsenal strategy is deemed preferred over the Liverpool and Manchester City strategy.

In section 2.4.4, the biases of Gambler's Fallacy and Hot-Hand Fallacy were presented, arguing that, according to the bettors, the performance of a football team is more dependent on the outcomes of past performances than they actually are. However, to what extent does the market efficiently price the probabilities of outcomes in such a situation? Does the bettors drive up the price of either outcome? - Or does the market participants suffering from Gambler's Fallacy and Hot-Hand Fallacy just cancel each other out in the aggregate market?

We have investigated this by implementing a simple condition; we only consider betting on matches where *one* of the participating teams have won the last *five* matches in a row. [14] In these matches, two opposite strategies are constructed: i.) bet on the team that have won their last five consecutive matches to continue winning (Hot-Hand Fallacy) and ii.) bet on the team that have won their last five consecutive matches to lose the match (Gambler's Fallacy). The results from implementing the two strategies across the five major leagues from the 2016-17 season until the 2023-24 season is illustrated in figure 2.4.

The two strategies depicted in figure 2.4 are very different. By always betting on a winning streak of five matches to continue, as a bettor suffering from Hot-Hand Fallacy, it generates a steady low-variance negatively-sloped linear curve. Bettors suffering from Hot-Hand Fallacy are seemingly highly exposed to betting on the expected, meaning that they will be the victim of the innate 5% commission from BetFair. Additionally, the odds will rarely be *high* in such instances. Out of 606 bets, only three of the won bets had odds above 4.00. 61.72% of the 606 bets were won with average odds 1.53.

Contrary to the low-variance negatively-sloped linear curve for the Hot-Hand Fallacy, the curve for the Gambler's Fallacy is high-variance and not close to being linear. 18.48% of odds were won with an average odds of 5.74. The Gambler's Fallacy is closely related to a long-shot strategy, which both are highly reliant on individual wins on bets with high odds. As briefly discussed in section 2.4.3, the

[14] It occurs that both participating teams in a match have won their last five matches, meaning we would be betting on both teams to win. As a consequence of this, such instances are removed from the investigation.

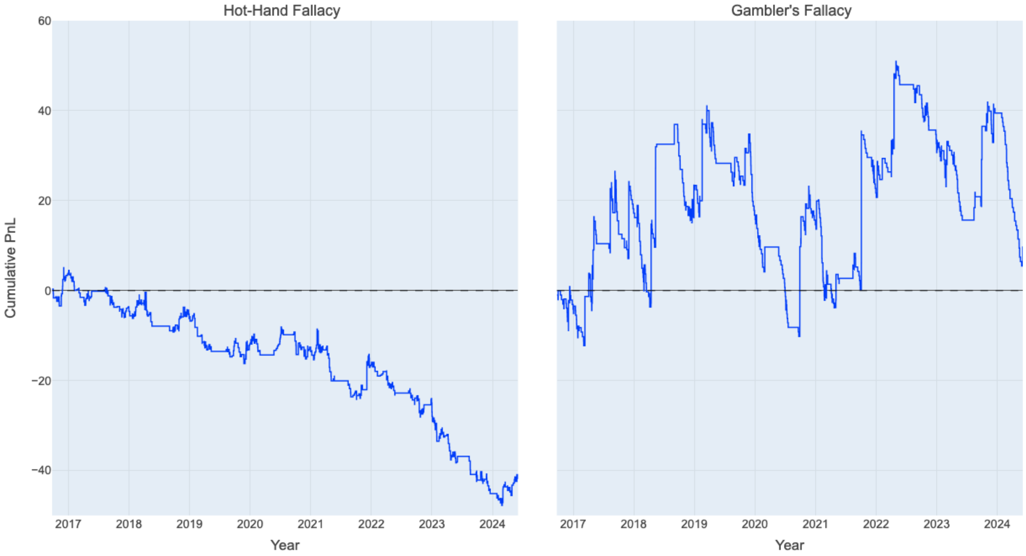


Figure 2.4: Hot-Hand Fallacy and Gambler's Fallacy: The figures illustrate the cumulative profits and losses (PnL) from the 2016-17 season to the 2023-24 season for two simple strategies: i.) bet a unit stake on the team that have won their last five consecutive matches to continue winning (Hot-Hand Fallacy) and ii.) bet a unit stake on the team that have won their last five consecutive matches to lose the match (Gambler's Fallacy). The figures are constructed based on all matches in the English Premier League, the Spanish La Liga, the German Bundesliga 1, the Italian Serie A and the French Ligue 1 from the 2016-17 season until the end of the 2023-24 season.

strategy is quite unpredictable (high-variance); while the strategy can only lose the stake of 1 per match in this setting, a handful of consecutive wins can turn a negative return into a significant positive return, however, what if these wins do not occur in the high frequency it has done throughout the historical data? E.g. in the 2022-23 season, out of 82 bets, only 15 bets were won, resulting in the returns going from 44.81 to 15.68. Equally, in the first half of 2024, only one bet was won out of 35 bets, resulting in the returns going from 38.45 to 9.82.

In section 2.4.4, a potential bias of emotional bettors were presented; the implied probability of victories for clubs with the largest fanbases might be overvalued, i.e. the odds are too low due to a irrationally high demand for such odds. We investigate this by always betting on a loss for all matches for the English team Manchester United and the Spanish teams Real Madrid and FC Barcelona, as they are among the clubs with the utmost largest fanbases globally. See figure 2.5.

While the strategies in figure 2.5 of always betting against Real Madrid and FC Barcelona generate positive returns most of the 8 year period, the variance is high, and even the most creative soul would find it problematic to explain it

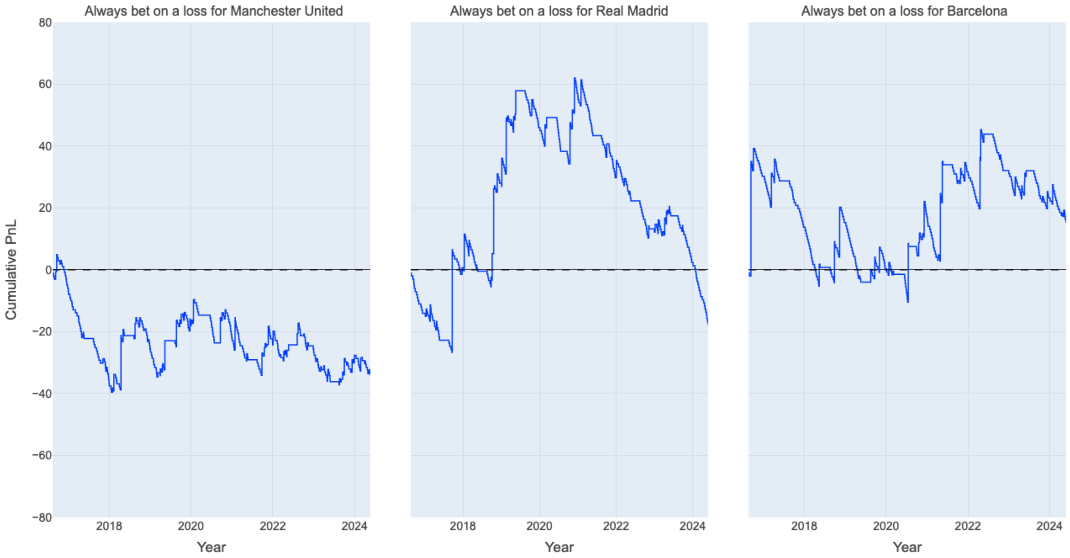


Figure 2.5: Always bet on a loss for each team: The figures illustrate the cumulative profits and losses (PnL) from 2016 to 2024 for a simple strategy: Always bet a unit stake on the team in question to lose - the teams being Manchester United, Real Madrid and FC Barcelona. The figures are constructed based on all matches in the English Premier League and the Spanish La Liga, which Manchester United, and Real Madrid and FC Barcelona have participated in from the 2016-17 season until the end of the 2023-24 season. The odds are the average supplied odds by the bet supplier *before* the match.

as a positively sloped linear curve. It is clear that the strategies concerning the two Spanish clubs are highly dependent on individual won bets. In fact, the Real Madrid strategy goes from a loss of -5.43 to a positive return of 57.88 in the span of almost eight months, by winning *only* 12 bets out of 32 wagers - six of which have odds above 5.00.

In this simple investigation, it seems that the betting market is somewhat efficient as the assumed large amounts of bets on their favorite club does not move the market prices to an extent that can be exploited.

3 Data and empirical method

3.1 Data & Variables

In this project, all datasets span from the summer of 2016 until May 1^{st} 2025, covering almost a decade of observations and approximately nine seasons of football matches. The relevance and characteristics of all datasets, as well as the exercise of combining the information from them, is presented in the below sections.

3.1.1 Bet Exchange Odds

The dataset of the bet exchange odds has come from Betfair by web-scraping oddsportal.[1] The data is from the summer 2016 until May 2025 for the English Premier League, the Spanish La Liga, the German Bundesliga 1, the Italian Serie A and the French Ligue 1, and the data consists of a datapoint for each individual decimal odds for all 1X2 outcomes of a match, hence, three rows per match. The odds displayed in *oddsportal* are the ones available immediately before the beginning of each match.

As Betfair has a 5% commission on profits from all won bets, we model the raw decimal odds and only subtract the 5% from the profits when computing the returns from a bet. Mathematically, the profits from a won bet are calculated based on the following equation:

$$\text{Net Positive Return} = (o_i - 1) \times b_i \times 0.95$$

, where o_i is the decimal odds and b_i is the stake. For lost bets, the loss is simply b_i.

In addition to the odds, the dataset provides the names of the home and away team, the timestamp for the match and the season, which allows us to combine odds data with the match data explained below.

[1] https://www.oddsportal.com

3.1.2 Match results

The data for match results have been scraped from FBRef. The data consists of a timestamp for the match, the season, the names of the home and away team as well as the number of goals that either team has scored in the match. We compute the results straight-forward:

$$\begin{aligned} &\text{if} \quad \text{HomeGoals} > \text{AwayGoals} &&\Rightarrow \quad \text{Home team wins} \\ &\text{if} \quad \text{HomeGoals} < \text{AwayGoals} &&\Rightarrow \quad \text{Away team wins} \\ &\text{if} \quad \text{HomeGoals} = \text{AwayGoals} &&\Rightarrow \quad \text{Draw} \end{aligned}$$

In practice, the odds and match result data is then combined into one dataset by grouping them by season and home and away team.

3.1.3 Predictive Features

In football, each match has a finite result and it is often said that a team is never better than their last result. However, it might be slightly more arbitrary than that; what if a team is fighting relegation - or if a team is fighting for the title? Contrarily, what if a team mathematically already have won the league? What if the manager is sacked if a match is lost? What if a team has a Champions League match in the mid-week? Numerous factors come into play, and they change on a weekly basis.

Some factors are challenging to implement as features, while others are more straight-forward. Below, we present the features that we have implemented (but not necessarily used in the final model).

Expected Goals (xG)

The Expected Goals data has been retrieved from FBRef by using the {soccerdata} library in Python.

Expected Goals (xG) is a statistical metric in football that quantify the expected amount of goals a team *should have* scored according to the calculated probability of a goal being scored in individual sequences. xG assigns a probability of each shot being a goal based on numerous parameters; type of shot (direct attempt on goal, header, volley, etc.), distance to goal, shot angle relative to the goal, presence of defender, goalkeeper position, etc. All else equal, the probability of scoring from the penalty spot is greater than from a shot on goal from 30 meters, hence the generated xG is greater.

The metric has become very valuable in football analyses, which provides rich information and can serve as a proxy for the *best* performing team in each match. If the actual number of goals for a team in a specific match exceed the expected goals, it indicates that the team is performing very well in finishing - and vice versa. Similarly, if a team had less actual goals relative to xG, it indicates ineffectiveness in finishing - or kudos to the goalkeeper.

The metric contributes in the ability to evaluate past non-result-based performances. Suppose a team has won their last several games. Are the victories deserved - or was it all luck? Now, suppose that the team has had roughly the same actual number of goals as xG, and the opposite team has had less than 1 xG in all those games. The conclusion would then be that the team is performing well, both offensively and defensively - the victories seem to be well deserved. The latter is useful information when predicting the probability of the team to win the next match. (Rathke, 2017)

In this project, xG is utilized in numerous ways: i.) difference in mean xGoals per game of the two teams throughout the season and in the past five matches, and ii.) mean difference between the actual goals and xGoals for each team throughout the season and in the past five matches.

Elo-ratings

The Elo-ratings data has been retrieved from ClubElo by using the {soccerdata} library in Python.

Elo-rating is a statistical ranking system that quantitatively classifies multiple entities according to the abilities of each entity. The higher the Elo-rating, the better the abilities. The Elo-rating then allows a ranking of n entities between 1 and n according to the relative skill levels (the Elo-rating) - 1 being the team/player/etc. with the weakest abilities and n being the strongest. Originally, the system was developed by chess master Arpad Elo in the 1960's as an alternative to the contemporaneous ranking system in chess, and the Elo-ranking system was implemented in chess 1970. The system is based on a logistic probability function of either entity winning the match, and after the given match, the Elo-rating is updated for both entities based on the difference between the actual and expected outcome of a match. The updated Elo-rating is given by the following equation:

$$
\begin{aligned}
R_i^* &= R_i + \gamma(S_{ij} - b(R_i - R_j)) \\
R_j^* &= R_j + \gamma(-S_{ij} - b(R_j - R_i))
\end{aligned}
\tag{3.1}
$$

, where b is an increasing function based on a constant scaling mechanism, meaning that the higher the difference in Elo-ratings between entity i and j, respectively R_i and R_j, the higher the change value. γ is the speed of adjustment, and S_{ij} is the result of the match; win, draw, loss, which takes the value $S_{ij} = 1$, $S_{ij} = 0.50$ and $S_{ij} = 0$, respectively. Therefore, intuitively, the greater the difference in Elo-ratings between the entities in a match $(R_i - R_j)$, the less impact on the change in Elo-rating from the result of the match. Contrarily, a match between two entities of similar Elo-rating have great impact on the Elo-rating after the match. (Düring et al., 2022)

The Elo-rating system is not new and has been subject to generalizations and improvements in research papers, including in (Düring et al., 2022) and Jabin et al. (2015). However, the fundamentals of the Elo-rating metric still persist in chess and is widely adapted in various other sports, including football.

One key advantage of the Elo-ratings is its simplicity and interpretability, and the ratings can directly be translated into win probability. As mentioned above, the ratings change continuously, taking form and momentum into consideration, and is therefore an obvious feature to include in the model of this present project.

As features in the XGBoost model, the Elo-rating is included as i.) the raw Elo-ranking for each team in each match, ii.) the difference between the Elo-rankings of the two teams, and iii.) as the rolling slope of the change in Elo-rating for each team.

Championship points

Based on the results of each match throughout the season, we assign a victory 3 points, a loss 0 points and a draw 1 point, i.e. we replicate the continuous standings in the league. The more points a team has, the better the team, and the better the chance for the title, and oppositely, the less points a team has, the greater the risk of relegation. It would be logical to believe that the team that need a victory in a match the most is one of the teams with either the most or the least points.

We have constructed numerous metrics based on championship points; i.) raw points in the season for each team (standings in the league table), ii.) mean points per match throughout the season and over the last five matches, and iii.) the points gap to the team with the most and the least points in the league.

Goals scored

A clever man once said that in order to win, one must score more goals than the opposition. Therefore, in addition to being utilized in combination with xG, we implement the mean goal difference (goals scored minus goals conceded) throughout the season and over the past five matches.

Odds

We have engineered features based on odds: The standard deviation of odds provided from *all* bookmaker platforms for both home team and away team victories, and the difference between i.) the odds spread between the odds for a home team and away team victory in bet exchanges and ii.) the odds spread between the mean home and away team victory odds across all bookmaker platforms.

Days since last game

While we only bet on games in the 5 biggest leagues of European football, there are also other tournaments taking place simultaneously, which can have an impact on a team's performance in the league. We can for example imagine a scenario, where one team is playing in both the league, the international Champions League and a domestic cup, while it's opposition only has the league to focus on. This would probably mean that the opposition has a higher-than-normal chance of winning, as the players on the other team is more fatigued from playing the previous game more recently and because they are more 'disinterested' in the main league, as they may deem a competition as the Champions League to be more important.

To try to capture these effects, we have included the number of days since the last game in any competition for both teams in each game.

3.2 Risk management

Risk management is fundamental for market participants in any investment or financial endeavor, and the sports betting market is no exception. However, unlike traditional financial markets, sports betting presents unique risks that require specialized approaches to capital allocation and assessment.

In cases of bankruptcy, investors are at risk of losing the entire invested amount when trading stocks, but it is a rarity. Contrarily, the bettor is at risk of losing the whole stake of each investment (bet), so the bettor must ensure that they can afford to lose the whole stake.

3.2.1 Kelly-Criterion

The Kelly-Criterion is a mathematical formula that quantifies the pre-defined optimal stake for individual bets, based on the objective of maximizing growth in the long run. Specifically, the deciding factor in regards to the size of the stake is a balance between risk and profit – obviously aiming at maximizing growth while minimizing the risk of ruin. The balancing of risk and profit is dependent on the odds of outcomes of each match, and the estimated probability of the same outcome. The Kelly Stake is calculated based on the general equation, in the context of betting: (Buchdahl, 2003, ch. 7)

$$k_i = \frac{e_i - 1}{o_i - 1} \tag{3.2}$$

, where k_i is the size of the Kelly Stake for the ith outcome, as a decimal proportion of the current bankroll. e_i is the *decimal edge* between the calculated probability of winning the bet and the market implied probability. o_i is the provided decimal odds of the outcome. In Buchdahl (2003, ch. 3), the *decimal edge* is given by the provided odds of the market divided by the bettor's calculated odds:

$$e_i = \frac{o_i}{\frac{1}{\hat{p}_i}}$$

, where $\hat{p}_i$ is the bettor's calculated probability. If the market odds are higher (implied probability is lower) than the calculated odds by the bettor, i.e. $e_i > 1$, the bet has a positive expected value for the bettor and is theoretically profitable in the long run – and therefore; all else equal, the greater the positive expected value for the bettor, the greater the Kelly stake.

In equation 3.2, it is clear that the Kelly Stake increases as the market odds decrease. Even though a bettor might find substantial positive expected returns in a bet when the odds are high, the probability of this happening is still low. However, if the expected returns remain positive, while the provided odds are low, the probability of the outcome happening is high, which leads to a higher Kelly Stake. This behavior of the Kelly Stake is based on the fundamental assumption that betting is a long-term activity - not a one-time occurrence.

As mentioned, the Kelly-Criterion quantifies the fraction of the *total current bankroll* that optimize the trade-off between potential profits and risk, which can lead to significant drawdowns. Suppose that the bettor sees a match with odds of a home team victory of 2.00, equating to an implied probability of 0.50, while the bettor finds the probability to be 0.80, equating to odds 1.25: Then $e_i = \frac{2.00}{1.25} = 1.60$ and $k_i = \frac{1.60-1}{2.00-1} = 0.60$, indicating that the bettor should have a stake 60% the size of her total current bankroll, according to the Kelly Stake. If that bet then loses, the drawdowns are significant.

Therefore, the **fractional Kelly-Criterion** is constructed to constraint the Kelly Stake from exceeding a pre-determined amount. The equation is given by:

$$k_i^f = c \times \frac{e_i - 1}{o_i - 1} \tag{3.3}$$

, where c is a fraction ($0 < c < 1$) to parameterize the tolerance between growth and risk. If $c = 0.50$, it is simple; only 50% of the total capital is available to the Kelly Stake, so the benefits of the principles of the Kelly-Criterion persists. (Buchdahl, 2003, ch. 7)

The practical limitations of the full and fractional Kelly-Criterion include the intrinsic assumption of validity of the bettor's ability to model the probability. Errors in estimating the probability of an outcome occurring can invalidate the decimal edge, potentially leading to drawdowns or suboptimal bet sizing.

Additionally, the exercise of identifying decimal edges against the market can be problematic, however, the identification is necessary for the Kelly Stake to be estimated. If there is no decimal edge, then $e_i < 1$ and $k_i < 0$.

3.2.2 Modern Portfolio Theory

Modern Portfolio Theory (MPT) is basically a way to think about constructing a portfolio of risky assets. The main idea behind MPT is that a rational investor always wants the highest return at a given level of risk - this means that if the investor were to accept more risk, she should also receive a higher return on the portfolio. These optimal portfolios, which maximize return at a given risk level is called *the efficient frontier*. (Hillier et al., 2012, p. 125)

Because the problem for the investor is as simple as it is, it is also possible to write a relatively simple optimization problem to describe it:

$$\max \mathbb{E}[P] - \gamma \text{Var}[P] \tag{3.4}$$

, where $\mathbb{E}[P]$ is the expected return of a portfolio, $\text{Var}[P]$ is the variance, indicating the level of risk for the portfolio, and γ is the risk-appetite of the investor.

It is normal to constrain the optimization problem in a few ways. If we denote the portfolio weights assigned to each of the n risky asset as w_i, two common constraints are:

$$\sum_{i=1}^{n} w_i = 1, \tag{3.5}$$

$$w_i \geq 0 \tag{3.6}$$

The portfolio we get from an optimization with the constraints in equation 3.5 is called the fully invested, long only portfolio - this is the exact constraint we will impose when running our portfolio optimization.

When working with MPT for equity markets, there are normally two approaches of estimation - ex-post estimation and ex-ante estimation. With ex-post estimation, the optimal portfolio for previous periods is used for the next period, while ex-ante estimation relies on the estimations of returns and covariance matrices. This can be a large task, when one is working with a lot of assets.

Utilizing MPT as a method for constructing a betting strategy is a little different. As a rule of thumb, assets are continuous on the stock market, i.e. if you were able to trade a stock in a previous period, you will be able to do it in the next period as well. On betting markets, on the other hand, every 'asset' (every outcome of a match) perishes after the underlying match has been played. This rules out ex-post portfolio optimization.

It can be argued, though, that ex-ante portfolio optimization is actually easier on betting markets than on equity markets. As we will see shortly, the only estimation we will have to make is the probability of each outcome of the match occurring. In contrast, the size of the problem explodes when doing this task for normal equities, as the number of assets increases, because the amount of moments between assets to estimate grows exponentially. (Brandt et al., 2009)

If we take equation 3.4 as our starting point, we need an equation for the expected profit and variance of the portfolio we are building. Much of what we are about to derive is heavily inspired by Hubáček et al. (2019), who use MPT to make an optimal portfolio for a betting strategy in a two-outcome game. We extend this to the three outcomes of a football game, and additionally, we differ slightly in our implementation of portfolio variance and in how we implement the procedure in practice.

For the betting market, the expected profit is quite simple, as we already know the payoff if we win or lose each bet. We just need to estimate the probability of the outcome happening:

$$P_i = \begin{cases} o_i b_i - b_i & \text{w.p. } \hat{p}_i \\ -b_i & \text{w.p. } 1 - \hat{p}_i \end{cases}$$

, which means that the expected profit is:

$$\begin{aligned} \mathbb{E}[P_i] &= \hat{p}_i(o_i b_i - b_i) + (1 - \hat{p}_i) - b_i \\ &= (\hat{p}_i o_i - 1)b_i \end{aligned} \tag{3.7}$$

, where $\mathbb{E}[P_i]$ is the expected profit of a single outcome with probability estimate $\hat{p}_i$, decimal odds from the betting exchange o_i and the wagered amount b_i.

The expected return for an entire portfolio with n games with 3 outcomes each is then:

$$\mathbb{E}[P] = \sum_{i=1}^{3n} \mathbb{E}[P_i] \tag{3.8}$$

For the expected risk of the portfolio, a very common measure is the variance. As we already have equation 3.7 for the expected value of a single outcome, it is quite simple to calculate the variance, as:

$$Var[P] = \mathbb{E}[P^2] - \mathbb{E}[P]^2$$

In our case:

$$\begin{aligned} \mathbb{E}[P_i^2] &= \hat{p}_i \left[(o_i - 1)b_i\right]^2 + (1 - \hat{p}_i)(-b_i)^2 \\ &= b_i^2 \left[\hat{p}_i(o_i - 1)^2 + (1 - \hat{p}_i)\right] \end{aligned}$$

and

$$\mathbb{E}[P_i]^2 = \left[(\hat{p}_i o_i - 1)b_i\right]^2$$

and therefore, we can write:

$$\begin{aligned} \mathrm{Var}[P_i] &= b_i^2 \left[\hat{p}_i(o_i - 1)^2 + (1 - \hat{p}_i)\right] - \left[(\hat{p}_i o_i - 1)b_i\right]^2 \\ &= b_i^2 o_i^2 \hat{p}_i(1 - \hat{p}_i) \end{aligned}$$

In order to calculate the variance for the entire portfolio, we can take two approaches. In Hubáček et al. (2019), they take the assumption that the bettor would never want to place a bet on more than one outcome in a single match. This would mean that the covariances of outcomes would become 0.

There are a few problems with this assumption, though. First of all, it is not clear how Hubáček et al. actually impose the assumption when optimizing their portfolio. Specifically, it is unclear whether they 'pick a winner' from each match, and then only run their optimizations on these predetermined winners, whether they run the optimization on all possible outcomes, and then pick the outcome for each match with the highest weight, or whether they use a third method altogether.

Secondly, we do not want to restrict ourselves from betting on several outcomes in each match. The reason they make this restriction in Hubáček et al. (2019) could be that they are working with a 2-way market (each match only has 2 possible outcomes), which makes it practically impossible to have positive expected values on both outcomes in a match. In the 1X2-market that we are working with, it is perfectly plausible that we predict a much lower probability for a home win than the betting exchange, yielding a positive expected value on both draw and away win.

For this reason, we extend the methodology of Hubáček et al. to also include the covariances between outcomes for single matches.

The first thing to note, is that the covariance between to non-independent random variables can be found as:

$$\mathrm{Cov}(P_i, P_j) = \mathbb{E}[P_iP_j] - \mathbb{E}(P_i)\mathbb{E}(P_j) \tag{3.9}$$

In this case, P_i and P_j is the profit from betting on any two of the three outcomes of a match. Therefore, the formula for $\mathbb{E}(P_i)$ and $\mathbb{E}(P_j)$ have already been derived in equation 3.7.

Deriving $\mathbb{E}[P_iP_j]$ is a little more complex, but we end up with:

$$\mathbb{E}[P_iP_j] = -b_ib_j\left[\hat{p}_i(o_i-1)+\hat{p}_j(o_j-1)\right] + p_kb_kb_j \tag{3.10}$$[2]

where k is the third outcome in a 1X2 bet, which means that:

$$\mathrm{Cov}(P_i, P_j) = -b_ib_j\left[\hat{p}_i(o_i-1)+\hat{p}_j(o_j-1)\right] + \hat{p}_kb_ib_j - \left[(\hat{p}_io_i-1)b_i\right]\left[(\hat{p}_jo_j-1)b_j\right]$$

From here, we can then calculate the variance of the entire portfolio as:

$$\mathrm{Var}[P] = \sum_{i=1}^{3n}\sum_{j=1}^{3n}\mathrm{Cov}[P_i, P_j] \tag{3.11}$$

[2] Refer to section A.1 for the mathematical derivation of the probability of the product of two non-independent discrete random variables.

Implementing MPT for football betting

Actually implementing these ideas in practice is not exactly trivial. We input equation 3.8 and equation 3.11 into equation 3.4, set all *b*s to 1, as it is literally the size of the bets we are looking to find[3], and maximixe the value of equation 3.4 using convex optimization with the {cvxpy} library in Python. The MPT optimization equation is therefore:

$$\max \quad \sum_{i=1}^{3n} \mathbb{E}[P_i] - \gamma \sum_{i=1}^{3n} \sum_{j=1}^{3n} \mathrm{Cov}[P_i, P_j] \tag{3.12}$$

Another important question is which matches should be included when optimizing the betting portfolio. We can obviously not optimize a portfolio of all the possible bets for a season, as that would introduce (a lot) of look-ahead bias, refer to section 4.3.1. Even optimizing a portfolio for a single matchday[4] would probably introduce look-ahead bias, as there is no guarantee that all data for the last match of the matchday is available at the game time of the first game.

For this reason, our 'portfolios' consist of all the matches played on the same game date, as we are fairly certain that all the data needed to compute predictions for the last match of a day is also available before the first match.

The problem with this procedure, is that there is not a fixed amount of matches on each game date. This means that some Thursday may only have 1 match to spread the bankroll over, while a standard Sunday may have 30+ games across the five leagues. Obviously, it does not make sense for the best outcomes on the Thursday to be given a much higher portfolio weight than an outcome on the Sunday with equivalent risk/reward ratio.

To solve this problem, we implement a simple scaling method, where we scale the capital available to each outcome with the number of games in that 'portfolio'. Of course, this only works as long as we use a fractional approach similar to the fractional Kelly-Criterion, when applying the portfolio weights to each outcome. This means that the fraction of the bankroll applied to each bet in the portfolio will be calculated as:

$$W_i^f = W_i^{MPT} \times c \times g$$

[3] Another way to look at this, is that we 'omit' the *b*s, and then what we are actually doing is to find the optimal size of the *b*s.

[4] In the following, we distinguish between a matchday and a game date. A matchday is the *round* of a football league, where each matchday consists of 10 games (for a 20-team league) scattered over several days. A game date on the other hand is simply a (calendar) date where 1 or more games are taking place.

, where W_f is the fraction of the current bankroll wagered on each outcome in the portfolio, W_{MPT} is the assigned weights from the portfolio optimization for each outcome, c $(0 < c < 1)$ is the maximum fraction of our bankroll wagered on a single outcome, and g is the number of games on that date. For each outcome, the wagered amount will then be: $b_i = w_i^f \times \text{Current Bankroll}$

3.3 Evaluation methods

When evaluating the betting strategy, the main objective for the strategy is to maximize the profits for the bettor. For this, *Compound annual growth rate* and *Total Returns* are used. Alongside the absolute profits, *Sharpe ratio*, *Max Drawdown* and *Win/Loss ratio* is employed as they account for the risk-adjusted returns, which ensures that profits are assessed relative to the level of risk.

Additionally, we include *Linearity* and *Win-percentage* for further comparison of models.

Compound Annual Growth Rate

We use Compound Annual Growth Rate (CAGR) as a measure of the returns accrues per year. It is implemented as a function of the cumulative PnL in the following equation:

$$CAGR = \left(1 + \frac{\text{Cumulative } PnL_i}{\text{Initial bankroll}}\right)^{\frac{365}{T}} - 1 \tag{3.13}$$

, where T the number of days the strategy has been generating results. The CAGR shows the percentage profit or loss per year of the strategy being in action.

Max Drawdown

The max drawdown is an alternative measure of the risk of a trading strategy, as it measures the largest drop between a peak and a trough. The larger the max drawdown is, the more risk the strategy possess. The max drawdown can not stand alone as a risk measure, though, as it does not measure how *often* a strategy suffers drawdowns.

The drawdown at time t is:

$$Drawdown_t = \frac{\max_t \text{Cumulative } PnL_i - \text{Cumulative } PnL_t}{\max_t \text{Cumulative } PnL_t}$$

, which means that the Max Drawdown is calculated as:

$$\mathit{Max\ Drawdown} = \max \mathit{Drawdown}_t \tag{3.14}$$

Sharpe ratio

The Sharpe ratio is a metric developed for investigating the risk-adjusted returns as an evaluation method in the financial markets. Specifically, the metric compares the excess returns to the risk-free rate with the volatility the investor faces by achieving greater profit. In the present project, the Sharpe ratio will be applied to assess the efficiency of a betting strategy, helping to distinguish the level of return – relative to the risk the bettor faces. In financial markets, the risk-free rate is often assumed to be various Treasury bonds, such as a 10-year Treasury bond.[5] However, the sports betting markets does not have an explicit non-zero risk-free rate, i.e. all profits are in excess of the risk-free rate, and hence; in the context of sports betting, the Sharpe ratio will simply be the profits relative to the risk:

$$Sharpe = \frac{\overline{\pi}}{\sigma} \tag{3.15}$$

, where $\overline{\pi}$ is the mean returns of all individual bets, and σ is the standard-deviation, which is never negative. Therefore, if the Sharpe ratio is positive, the mean returns are positive, i.e. the investigated strategy generates profit. Intuitively, the greater the Sharpe ratio, the greater the profits relative to the risk. Consequently, maximizing the Sharpe ratio involves both maximizing the mean returns – and minimizing the risk. (Mcmillan, 2018, ch. 3)

In order to make the Sharpe ratio more standardized, we annualize it by multiplying the simple per bet Sharpe ratio with the square root of the average number of bets per year, denoted by n_y. This means that the formula for the annualized Sharpe ratio, which we will be using in the results section, is:

$$Sharpe_{\text{ann}} = \frac{\overline{\pi}}{\sigma} \times \sqrt{n_y} \tag{3.16}$$

As the Sharpe ratio quantifies the earnings of the bettor's strategy per unit of risk, it is a useful benchmark for comparing betting strategies. In sports betting, the risk of extensive drawdowns is innately significant, as the bettor losses the whole stake when losing a bet, and, assuming the mean returns remain constant, a higher Sharpe ratio indicates reduced exposure to such drawdowns. However, due to the mentioned structure of sports betting, both the profits and especially losses from

[5] The general equation of the Sharpe ratio: $Sharpe = \frac{\overline{\pi}-r_f}{\sigma}$, where r_f is the risk-free rate.

individual bets are large. This will increase the standard-deviation, and it will not differentiate whether the increased risk is due to a losing or winning streak. Fortunately, this is the case for all strategies in sports betting, so comparability remains.

Suppose a bettor implements a favorite strategy (mentioned in section 2.4.3), where the bettor only bets on teams with winning odds of less than, say, 1.10; One single lost bet will increase the standard-deviation massively, and perhaps, the profit might evaporate, as the bettor can only afford to lose one bet from ten won bets with odds 1.10. In this scenario, the mean return will fall and standard deviation will increase, lowering the Sharpe ratio substantially.

Though, in other strategies, individual losses and wins do *not* have these extensive consequences. However, for a bettor that bases their strategy on the Sharpe ratio, the justification of a less conservative strategy (relative to the favorite strategy), the increased risk must necessarily be followed by an increase in profits.

Linearity

Linearity is more of a home-brewed measure of a strategy's performance. When looking at a strategy, we want it to be as close to a linear line as possible, as this would mean a strategy of steady return and low volatility.

To assess how closely a strategy's bankroll evolves along a straight line, we run a linear regression on the current bankroll with the number of days since the strategy's inception as the only predictor:

$$Bankroll_t = \alpha + \beta \, t_i$$

, where $Bankroll_t$ is the bankroll at time t, α is the regression's intercept, β is the slope (increase in bankroll per day after the strategy's inception), and t is the number of days since the strategy's inception.

The linearity is then measured as this regressions' R^2-value.

The closer this linearity measure is to 1, the better we deem the strategy to be. Together with other measures of profit and volatility it contributes to give a thorough view of a strategy's performance.

Win/Loss ratio

The win/loss ratio is a metric that finds the relative difference between the daily returns, i.e. the percentage increase/decrease in the bankroll on each game day, of bets that have been won and those that have been lost. Mathematically:

$$\text{Win/Loss ratio} = \frac{\frac{\sum^{N_w} y_{i,w}}{N_w}}{\frac{\sum^{N_l} y_{j,l}}{N_l}} = \frac{\mu_W}{\mu_L} \tag{3.17}$$

, where $y_{i,w}$ is the returns on all won bets on game day i and $y_{j,l}$ is the returns on all lost bets on game day j. N_w and N_l are the number of game days, where the bettor have experienced won and lost bets, respectively. Obviously, the bettor will prefer a higher Win/Loss ratio.

Win-percentage

Win-percentage (WinPct) is the percentage of all won bets out of the total number of bets. Therefore:

$$\text{WinPct} = \frac{\text{Number of won bets}}{\text{Total number of entered bets}} \tag{3.18}$$

Total Return

Total Return is a measure for the total percentage *change* in the bankroll from the first day of betting until the last day of betting:

$$\text{Total Return} = \frac{\text{Final bankroll}}{\text{Initial bankroll}} - 1 \tag{3.19}$$

4 Modeling

4.1 XGBoost

The *eXtreme Gradient Boosting* (XGBoost) model was introduced by Chen & Guestrin in 2016, and it is a scalable and efficient implementation of the *gradient boosting framework*, which has achieved great traction and adaptation in machine learning disciplines. Contrary to traditional econometric models, which prioritize statistical inference and interpretability, the XGBoost model focuses on the out-of-sample prediction performance. As a consequence of this, XGBoost is often referred to as *"black box technology"*.[1]

At its core, the XGBoost is a combined model of numerous decision tree models that are *weak* in its predictive ability, hence its name: *weak learners*. Decision trees are prone to overfit the training data, and the more noisy the data is, the more overfit the decision trees are likely to be. Therefore, the decision trees in the XGBoost framework are trained to explain the data as accurately as possible, obviously, however; the decision trees are heavily penalized when nodes split in order to prevent being overfit. In other words; each individual decision tree must only split its branches an additional time, if the explanatory power increases significantly from doing so. The combination of weak learners in order to generate one with high predictive accuracy is known as *Boosting*. (Lopez de Prado, 2018, ch. 6)

A **decision tree** is a non-parametric model, that is constructed like a tree. It recursively narrows the *feature space* into smaller spaces that increasingly becomes closer to the prediction of the target variable. These recursive steps are called nodes; at each node, where a branch becomes two, the algorithmic model selects the feature and threshold that best splits the data to minimize the pre-determined loss

[1] A device, model or system, that is said to be a black box, produces useful information as any other model - but with little or no information about the internal workings.

function. One can view this process as a series of *if-then* rules that together form a tree-like structure, and where each leaf is a prediction of the target variable. (James et al., 2023, ch. 8)

Suppose a decision tree: The first split of the tree is the *if-then* rule that best partitions the dataset between low-and-high probability of winning. Suppose that this rule is: victories in the previous five matches > 3; the teams with a lot of recent wins are much more likely to win their next game than a team with fewer recent wins. The dataset is now split into two parts, which the model will once again try to split in a way where the difference between the two splits is maximized. The more times we split the tree, the harder the thresholds for further splitting are to fulfill, which means that at some point, it makes no statistical sense to continue splitting the data. According to Lopez de Prado (2018, ch. 9), classification decision trees usually split based on a classic *log-loss function* (also known as the cross-entropy loss function)[2]:

$$L[Y,P] = -\log[Prob[Y|P]] = -N^{-1}\sum_{n=0}^{N-1}\sum_{k=0}^{K-1} y_{n,k}\log[p_{n,k}] \tag{4.1}$$

, where Y is the binary indicator matrix of K *true* possible classes of the target variable, and P is the probability distribution of all N predictions of the model. Therefore, the individual prediction $y_{n,k} = 1$ only when the n^{th} datapoint truly belongs to the k^{th} class, regardless of the prediction. p_{nk} is the probability of datapoint n being classified in the true class k, and consequently, the log-loss equation *only* evaluates the probability corresponding to the true class prediction $y_{n,k} = 1$ and the model's probability thereof p_{nk}. Equation 4.1 is a minimization problem, and intuitively, the greater the probability p of predicting correctly, when the prediction is correct, the smaller the punishment (hence the name log-loss). Note, the log-loss is calculated based on the probability assigned to the true class, and therefore an incorrect prediction implies a low probability for the true class, leading to a higher punishment. It is therefore important to mention that the objective is not *predict correctly*, it is rather *minimize the surprise of the true outcome*.

In practice, when a decision tree is in action, it considers different features and thresholds for each split. The tree considers dividing the data into at least two groups and the split is evaluated based on the log-loss within each hypothetical group. The split that leads to the greatest reduction in log-loss is then compared to the log-loss of the parent node, i.e. does it even make sense to make the split?

[2] A frequent alternative to the implementation of the log-loss function is the *Gini index*

The objective of each split is to create a child node that finds the true classes more predictable.

This process is iterated until the log-loss is *not* reduced from splitting the data once more. At this node, the data is split into regions where predictions are most accurate, according to the designated objective function; the log-loss. (James et al., 2023, ch. 8)

As briefly mentioned, decision trees are oftentimes overfit, i.e. decision trees tend to explain the training data *too well*, and is therefore incapable of explaining new information that is unsimilar to the training data. To combat overfitting, **Boosting** is used, which is an ensemble technique that combine numerous weak leaners – or shallow trees. The aggregation of weak learners forms a strong predictive model which is capable of explaining the underlying tendencies in the data, rather than the exact tendencies the training data shows.

The key idea is to sequentially fit new trees that correct the residuals of the previous trees by fitting a decision tree as a function of the residuals – not the outcome variable. Therefore, unlike *bagging*[3], boosting is highly dependent on the trees that have already been constructed. (James et al., 2023, ch. 8)

Simple boosting has three hyperparameters to tune: the number of trees B, the learning rate λ and number of splits in each tree d. In boosting, if B is *too* large, the aggregated prediction model tend to be overfit, but often, B is estimated using cross-validation. While the boosting framework *learns slowly* by construction, relative to the *ordinary* decision tree, λ, which is always a positive number, can be changed. Higher λ means that the model learns faster but faces greater risk of overfitting. Contrarily, lower λ generates a slower and more cautious model, which necessitates a large value of B. The number of splits in each tree d is often equal to the number of explanatory variables, as d splits can involve no more than d variables. (James et al., 2023, ch. 8)

Consider the equation:

$$\hat{f}_2(x) = \hat{f}_1(x) + \lambda \hat{f}^b(x) \tag{4.2}$$

, where $\hat{f}_2(x)$ is the *updated* aggregated decision tree, based on the previous tree $\hat{f}_1(x)$ and the residual tree (new learner) $\lambda \hat{f}^b(x)$ and the learning rate $\lambda \in [0, 1]$.

[3] Bagging builds multiple independent models, which are constructed based on bootstrapped subsets of the data. The predictions of the models are then averaged, and the aggregated residual minimized. Therefore, while boosting works sequentially and weights its learning, bagging constructs multiple models simultaneously. (James et al., 2023, ch. 8)

Therefore, the "new" residuals are:

$$r_2 = r_1 - \lambda \hat{f}^b(x) \tag{4.3}$$

Each tree, or new learner, simply models the residuals of the past trees, and therefore the *final* boosted model is given by:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{4.4}$$

, where the boosted model is given by the sum of all B individual learners b, on the basis of the optimal value of λ and the number of splits d.

In practice, the above-presented procedure of boosting reviews and empathizes on improving samples of the training data, which is poorly explained, and therefore; has poor predicting ability. This is done through a sequential adjustments of weights, increasing the influence of poor predictors.

The *eXtreme Gradient Boosting* (XGBoost) extends the technique of boosting, denoted *Gradient Tree Boosting* (GTB), hence its name. In contrast, GTB reframes the process of boosting as an optimization problem. Models are still added sequentially with the objective of minimizing the pre-determined loss function, however, with the knowledge of the negative gradient taken into consideration. Each new learner is fitted to the *negative gradient* of the loss function, i.e. with respect to the slope of the loss function. Therefore, new learners are fitted more optimally and in a more flexible manner, as the principles of gradient descent help guide the learning process of new learners. (Lopez de Prado, 2018, ch. 22)

In addition to incorporating the classical GTB framework, XGBoost introduces several algorithmic enhancements and systematic optimizations[4], making it a widely adopted real-world application when forecasting data-rich variables. These applications are all constructed on the foundation of the decision tree model, but it introduces a regularized objective function and a second-order optimization scheme, improving its predictive accuracy and generalization capability of complex information. (Chen & Guestrin, 2016)

The first-mentioned regularized objective function is given by:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4.5}$$

[4] Regularized Learning Objectives, Shrinkage and Column Subsampling, sparsity-aware algorithm. Block Structure for Parallel Learning, Cache-aware Access Patterns, Weighted Quantile Sketch, etc. (Chen & Guestrin, 2016)

, where $l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))$ is the differentiable *convex loss function* at the t^{th} iteration and the i^{th} instance. Specifically, the first term measures the distance between the prediction $\hat{y}_i$ and the target y_i, while the second term $f_t(x_t)$ is the decision tree that has removed (or partially removed) the errors at iteration t of the feature x. Additionally, $\Omega(f_t)$ is the *regularization component*, which penalizes the complexity of the model in respect to another tree being introduced, represented by f_t.

The regularization component is given by the *regularization equation*:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4.6}$$

, where T is the number of leaves in the tree f_t and w_j is the weight of each leaf j in the tree. Hence, γ and λ are *regularization parameters*, that penalize each additional leaf and high weights on individual leafs, respectively. The objective is to minimize the risk of overfitting.

By improving the learning process of new learners in an XGBoost model, i.e. to efficiently minimize the objective function, equation 4.5, the second-order derivative thereof is employed. This second-order approximation is given by:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{4.7}$$

, where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are the first and second-order gradient statistics of the loss function, respectively. The gradient g_i indicates the direction of improvement in minimizing the error in prediction accuracy when adding a new tree, while the Hessian h_i explains the curvature of the loss function. This allows the algorithm to adjust the direction and magnitude appropriately to minimize the loss function. The result of this method is faster convergence to the model with *best* specifications, than that of *ordinary* boosting, which reduces the needed iterations to reach a satisfactory model. (Chen & Guestrin, 2016)

In addition to the above, XGBoost handles various other computationally intensive and econometric exercises cleverly. For example, when data are missing, *sparsity-aware algorithms* are implemented, which learn the optimal default direction for missing data values. In cases of missing data, the node directs the prediction in this default direction. Models such as XGBoost generally require substantial amounts of computational power, and in order to minimize this, XGBoost stores *old* learners efficiently in compressed column blocks, denoted a *Column Block for Parallel Learning*. (Chen & Guestrin, 2016)

Although the risk of constructing an overfit model is reduced significantly when implementing the XGBoost, the risk is not eliminated. Refer to section 4.3.

4.2 Market decorrelation through custom loss function

It can be argued that it is naive and perhaps even cocky to believe that we can systematically beat the betting market or a bookmaker by outputting a superior prediction of the outcome of a match. As explored earlier, inefficiencies can occur for different reasons, but still it makes sense to explore other options to beat the market than simply by 'being better'.

In Hubáček et al. (2019) and Hubáček & Šír (2023) a quite powerful idea is presented; you do not need an especially good predictive model to beat the market, your predictions just need to be suitably decorrelated with that of the market. In this context, the betting markets are a natural case, as we always have access to the market forecast; the odds on the betting exchange.

The hypothesis that an outcome prediction needs to be decorrelated with the market forecast to make money is almost self-explanatory in the context of a betting market. If we always make the exact same prediction as the market, we will in the long run simply lose the 5% commission that the exchange takes on profits, as explained in section 2.1.

Hubáček et al. explores several ways to decorrelate their outcome prediction with that of the market, including adding sample weights to each prediction based on the odds on that outcome.[5] They also present a much more elegant method, which is to add a penalization term to their model's loss function, to penalize similarity to the market forecast.

Even though their problem is a binary classification problem[6], the starting point for their loss function is the mean squared prediction errors (MSE), which is the standard loss for a regression problem. In its standard form, it looks like:

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{4.8}$$

[5] The argument here is that the higher the odds, the more important the datapoint is, as that is where the potential profits are the greatest.

[6] They want to predict whether a team wins or loses, so $y_i \in \{0,1\}$

, where N is the number of predictions, $\hat{y}_i$ is the prediction for the ith observation and $y_i \in \{0, 1\}$ is the actual value for the ith observation.

If we want to decorrelate our forecast with the 'market forecast', which in this case is simply the decimal odds, we penalize predictions that are similar to that of the implied probability of the betting exchange. We do this by adding an extra term to the above loss function:

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 - c \cdot (\hat{y}_i - 1/o_i)^2 \tag{4.9}$$

, where $c \in [0, 1]$ is a penalty parameter determining the relative weight of the decorrelation term compared to the MSE term and o_i is the market odds.

The objective of a machine learning algorithm is to minimize the loss function, which in this project is the argument for subtracting the decorrelation term from the original loss function. The more dissimilar our probability prediction $\hat{y}_i$ and the implied probability from the market odds $\frac{1}{o_i}$ are, the larger the decorrelation term is, and the smaller the value of the loss function will be.

As mentioned, the MSE is usually used for regression problems, as it does not provide any restrictions on the value of $\hat{y}_i$. This means that we can end up with probability estimates exceeding 1 or under 0, which, of course, does not make sense when working with a classification problem. (Wooldridge, 2012, p. 584). This means that we would have to deal with the problem of how to map these probabilities after the fact; should a probability of 1.25 simply be truncated to 1 or should all probabilities be normalized in order to preserve the magnitudes of differences between the probabilities?

Additionally, it can be shown with a simple example that the MSE in equation 4.8 does not do us any favors when doing our optimization for a classification problem, compared to the log-loss function, which we will explain shortly. If we predict a probability of 0.01 of an outcome to have the label 1, but the actual label of the outcome *is* 1, the losses would be:

$$\text{MSE} = (0.01 - 1)^2 = 0.98$$
$$\text{Log-Loss} = (-)1 \cdot \log(0.01) = 4.81$$

The difference comes down to the fact that the log-loss penalizes large deviations like this a lot more than the MSE, meaning that the MSE will have a harder (and slower) time coming to a conclusion when trying to optimize the model, which in the end could give us an inferior model.

In order to avoid the problems that come with using a regression-based loss function for a classification problem, we have chosen to extend their application to the classic log-loss (also known as logistic loss or cross-entropy loss), which, as mentioned in section 4.1, is the standard loss function for binary classification problems in the XGBoost algorithm.

A more specific form of equation 4.1 is the log-loss for a classification problem with two possible outcomes, which is defined as:

$$-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\hat{p}_i) + (1-y_i)\cdot \log(1-\hat{p}_i)\right] \tag{4.10}$$

, where $\hat{p}_i$ is the logistic transformation of the raw probability output, $\hat{y}$, for observation i:

$$\hat{p}_i = \frac{1}{1+\mathrm{e}^{-\hat{y}_i}} \tag{4.11}$$

In order to then decorrelate our forecast with the implied probability from the market odds, we add the same penalty term as in equation 4.9:

$$-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\hat{p}_i) + (1-y_i)\cdot \log(1-\hat{p}_i) + c\cdot(\hat{p}_i - 1/o_i)^2\right] \tag{4.12}$$

In contrast to equation 4.9, we add the decorrelation term to the original loss function. This is because the log-loss is already negated (with the $-\frac{1}{N}$-term).

Therefore, in this model, we simultaneously reward the model for predicting correctly - but also for providing probabilities that *deviate* from the probabilities implied by the odds. The rationale behind this approach is that simply attempting to imitate the implied probabilities, or at least trying to mirror the probabilities to a certain degree, would, over time, lead to a loss equivalent to the overround, mentioned in section 2.1. By partially focusing on deviations from the market odds rather than directly competing the ability of the market to calculate probabilities, fundamentally, the model is attempting to find various possible patterns that the market does not find.

4.2.1 Implementing the loss function in XGBoost

As already explained, the market decorrelated loss function in equation 4.12 only works for a binary classification problem, i.e. $Y \in \{0,1\}$. Football matches are unfortunately not quite that simple to bet on, as there are three possible outcomes; home win, draw and away win. As mentioned, this is often called 1X2-betting, so we can denote it $Y \in \{1, X, 2\}$.

The problem we face here is that a so-called multi-class classification loss function, i.e. a loss function for a classification problem where the true label can take more than 2 values, is significantly more complex than ditto for a binary classification problem. As explained in section 4.1, we have to obtain the gradient and the hessian of the loss function for the XGBoost algorithm to work.

When calculating the gradient and hessian of a loss function for a binary classification problem, e.g., the loss in equation 4.12, we only have to calculate the derivative and second derivative w.r.t. the prediction $\hat{p}_i$, which would be a vector in both cases. As can be seen in the example in xgboost developers (2022), we need to calculate the partial derivative and second partial derivative w.r.t. to all possible outcome classes in a multi-class classification problem. While this is certainly possible to implement, it would require much more time and debugging. Additionally, the implementation would be quite slow in a programming language like Python, which we are using, as we would need nested for-loops to make the computation, as can be seen in the example in xgboost developers (2022).

For this reason, we take a different approach, where we simply extend a binary algorithm to a multiclass problem using the one-vs-rest method. This means that we train 3 separate models to predict whether one class is true or not. Recall that our possible outcomes are: $Y \in \{1, X, 2\}$, giving us three models.

$$\begin{aligned} f_1(x) &\approx P(Y = 1), \\ f_X(x) &\approx P(Y = X), \\ f_2(x) &\approx P(Y = 2). \end{aligned}$$

Although this gives a probability of each of the three outcomes happening, there is no guarantee that the outcomes sum to 1. To alleviate this, we simply normalize the probabilities to sum to 1, by dividing each individual probability with the sum of the probabilities for all classes:

$$\hat{f}_c(x) \;=\; \frac{f_c(x)}{\sum f_c(x)}, \quad c \in \{1, X, 2\}.$$

4.3 Avoiding backtest overfitting

Backtesting can be defined as a technique to assess how a trading strategy *would have* performed, should it have been run over a past period of time. As with any research in social sciences, backtesting does not serve as a typical 'experiment' as in physics: even in its most flawless form, a backtest does not prove anything in itself. (Lopez de Prado, 2018, p. 151)

In spite of that, a backtest is the best tool at a quantitative researcher's disposal when assessing the quality of a trading strategy. After all, it would be foolish to wager big money on a trading strategy which does not at least have a solid historical track-record. This fact makes it all the more important that we do not *overfit* our backtest, which is defined as fitting the strategy to random historical fluctuations, which will not reoccur in the future, resulting in poor performance on new, unseen data. According to Lopez de Prado (2018), backtest overfitting could even be considered scientific fraud, as it is argued that conducting an experiment over and over on the same data, will eventually lead to a false discovery.

While backtest overfitting is arguably the greatest problem in all mathematical finance, unfortunately, there is currently no definitive solution to overfitting. Especially for complicated machine learning models that are constructed to capture complex patterns, backtest overfitting is a constant threat to out-of-sample predictive ability. (Lopez de Prado, 2018)

Instead, researchers must employ other techniques such as cross-validation, out-of-sample testing and necessarily critical interpretation of in-sample results to guard against overfitting and ensure that a model's predictive power extends beyond the data it was trained on.

In order to prevent this problem, and, in the eyes of Lopez de Prado, not commit scientific fraud, we employ a few central techniques to avoid overfitting our backtest. Additionally, we will in our results section comment on whether and to what degree we may have overfitted our backtest. Refer to section 5.5

4.3.1 Look-Ahead Bias

Perhaps the most important bias to eliminate is the so-called look-ahead bias, where future information is used to make a decision on a past datapoint.

The first step to eliminate this problem is in the data engineering step, where we have to make sure that all variables only use past information. While this perhaps sounds very simple and obvious, it is a crucial step to always remind one-self of. For example, when computing the average points in previous matches, it is important *not* to include the *current* match. When training a model based on information that is not available at the time of forecasting, it is called *leakage*. As a consequence of this, the model can potentially be evaluated under unrealistic conditions, which can lead to inflated estimates of the model's ability that does not align with real-world prediction scenarios. After all, it is easier to predict the outcome of a match, if one knows the change in league points for each team in the match. Such leakages can occur with seemingly strict temporal separation, so, according to Lopez de Prado (2018, ch. 12), a carefully implemented temporal control is necessary to obtain robust backtesting.

In Lopez de Prado (2018, ch. 7), the exclusion of datapoints with informational overlap when splitting the data between training and testing data is proposed as a solution to leakage, which we will implement in our model. A methodology such as purging is especially necessary to implement when doing *K-Fold Cross-Validation* (purged K-Fold CV), however, in the present project, we implement the *walk-forward method*, ensuring a chronological order in the training and testing data: The training data will *always* come before the testing data.

Walk-forward is a common choice for cross-validation in quantitative finance strategy. The walk-forward method has a simple approach; the model is trained on past data, and the model is tested on future data. This is done for a subset of the data, and then the window is moved forwards (in the direction of time) and the process is repeated (re-training). Clearly, the interpretation of conducting a walk-forward backtest has a clear and chronological interpretation, however, the non-customized walk-forward method is prone to overfit the training data, as there is a historical path of testing, which can be repeated again and again, until some pattern is found. The found pattern could potentially be based on tendencies that is not rooted in the real world. (Lopez de Prado, 2018, ch. 11 & 12)

Purging is the introduction of a time period between the training period ends and the testing period starts, i.e. we deliberately skip a pre-determined (short) amount of data before we start our testing period, which acts as a buffer. By doing this, we are *extra* careful to make sure there is no leakage of data in the near future. (Lopez de Prado, 2018, ch. 7)

Mathematically, suppose two concurrent datapoints Y_i and Y_j are in the training data and the testing data, respectively. In the model, the probability computation of outcome Y_i is based on features ranging from time $t_{i,0}$ to $t_{i,1}$, and equally $Y_j = [[t_{j,0}, t_{j,1}]]$. Leakages occur if one of the below conditions are met: Lopez de Prado (2018, ch. 7)

- $t_{j,0} \leq \mathrm{t}_{i,0} \leq \mathrm{t}_{j,1}$ (Testing window starts within Training window)
- $t_{j,0} \leq \mathrm{t}_{i,1} \leq \mathrm{t}_{j,1}$ (Training window starts within Testing window)
- $t_{i,0} \leq \mathrm{t}_{j,0} \leq \mathrm{t}_{j,1} \leq \mathrm{t}_{i,1}$ (Testing window ends within Training window)

The length of the purge period must be decided based on thorough consideration in regards to the characteristics of the data at hand. In the present project, the computed purge is *one day*, which means that we, for example, reoptimize the model on April 1^{st} based on data until and including March 31^{st} and the testing data starts April 2^{nd}. This ensures that there is no leakage between the last day of the training data and the first day of the testing data.

Apart from a purge period of 1 day, we are using a *rolling* walk-forward method, where the training window is a minimum of half a year and then stays fixed at 1 year and a testing set size of a month. This means, that we train our model on the games in the previous year after which we test the model on the games in the next month. Then this month gets included in the training dataset, while the 'oldest' month gets removed, to keep the training set fixed at 1 year. This procedure is repeated iteratively until the entire dataset has been exhausted. This procedure is visualized in figure 4.1 for the 2021-22 season and onward, to make it fit nicely on the plot. As is clear from the plot, each training and test set does not necessarily have the same size due to breaks in the football calendar.

4.3.2 Data-Snooping

Closely related to the above-mentioned Look-Ahead Bias, *data-snooping* occurs, in our case, when datapoints are both used for constructing betting rules *and* testing those betting rules, which violates the independence of samples. Specifically, when the strategic bettor is constructing betting rules by identifying patterns, tendencies and relationships in the historical data, denoted *data-mining*, data-snooping occurs when historical data is used repeatedly and simultaneously to both identify and test the validity of the identified patterns. The problems arise as the data-mining process is repeated, and non-significant patterns are identified, i.e. spurious relationships. This can lead to the researcher concluding significant positive returns that in reality are based on random chance.

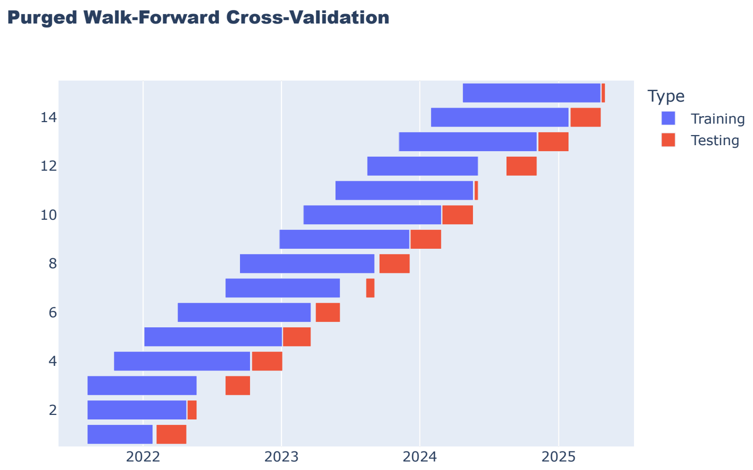


Figure 4.1: Our rolling, purged walk-forward procedure since the 2021-22 season. Because of breaks in the football calendar year has several breaks, which is why each training and test set does not necessarily have the same size each time.

In practice, after multiple iterations of in-sample tests of various betting rules, the seemingly significant betting rules are carried forward into the subsequent testing and forecasting phases, while betting rules that did not provide a positive return is discarded. In essence, in the long term, the researcher has a model constructed on betting strategies that performed well in-sample, however, some of which might be biased and insignificant and generated purely by chance, denoted *inter-generational data-snooping*.

While the risk of concluding spurious relations are prominent when working with poor amounts of data, it is important to note that the risk of identifying and going forward with spurious relationships persists as the amounts of data increase. In financial econometrics, where datasets oftentimes are very large and granular, the use of data-mining and data-snooping are widespread. However, currently there is no definitive solution for this, instead the researcher must find variables based on financial and economic theory and implement a clear methodology for i.) constructing betting strategies and the model and ii.) the validation thereof. (Brooks, 2019, ch. 4)

First, all explanatory variables included in the model are selected on the basis of theoretical justification as factors that is capable of painting the picture of a team in terms of ability, form and reciprocal differences between the two teams in each match. This will reduce the risk of including variables that *might* or *might not* have non-spurious explanatory power. Had we included weather or week-day of each match, it is possible that certain teams have a relative advantage when matches are played on Wednesdays in the rain, but it might just be a coincidence, i.e. a spurious relationship between weather or week-day and winning probability.

Second, in order to combat the risk of introducing spurious relationships in the forecast, we implement a validation step (a hold-out set of the data) before performing the out-of-sample forecast. Therefore; the in-sample training data is used exclusively for training and selecting betting strategies, the betting strategies are then tested and validated in our validation step in the discrete time-period following the training data, and *if* the validation step generates satisfactory results, the out-of-sample forecast will be performed on the testing data, which *we only look at once*.

The validation step serves as an out-of-sample barrier to evaluate the robustness of the betting strategies found in the training data *before* the actual out-of-sample forecast that generates the actual results. The objective of the additional layer is to put attention to betting strategies that only perform well in-sample. We then adjust the betting strategies and the model until we are satisfied with the validation before finally forecasting and comparing to the testing data. (James et al., 2023, ch. 5)

We use the seasons from 2016-17 to 2022-2023 as the in-sample dataset, season 2023-2024 as our validation set and season 2024-2025 as the out-of-sample dataset. Additionally, we exclude season 2020-2021 both when training and evaluating our models and strategies, because the underlying statistical properties of that season was so heavily influenced by the Covid-19 pandemic - we touch more upon this in section 5.2.

5 Results & Discussion

5.1 How well does the betting exchange predict match outcomes?

The commission fee of 5% on profits, which Betfair charge on profits from a wager in sports betting is considerably more than the transaction fees in e.g. stock trading. Developing a model capable of outperforming a simple buy-and-hold strategy in the stock market presents a significant challenge - a challenge that attracted substantial attention from academics and researchers within financial econometrics. Suppose the transaction costs in the stock market increase to the levels of the betting market; a very difficult task becomes even more difficult.

One can argue that the strategic bettor's probability model must outperform the concerned market to a higher degree, than a model in the realm of stock trading, for example. In other words, the modeling-wise burden of proof for profitability in order to enter a wager/trade is significantly heavier in the betting markets. While both fields of price forecasting has its foundation in inefficiencies and forecasting accuracy, the modeled probability threshold for sustainable profitability in sports betting are higher than in traditional financial markets. Contrarily, it is reasonable to conclude that the frequency of researchers attempting to exploit inefficiencies in the financial markets are significantly higher than researchers attempting to exploit inefficiencies in sports betting.

Throughout the entire dataset that has been processed in this project, the market accuracy in the 1X2 market is 54.14%. Therefore, in more than half of all possible matches, the favorite team to win, according to the market odds on Betfair, has won the match. This high market prediction accuracy reflects the level of efficiency in the aggregated market where public and perhaps private information is incorporated in the market mechanism of supply and demand. However, this also raises the question of whether it is necessary to have proprietary information that the

Table 5.1: Descriptive statistics comparing pre- and post-Covid subsets of the data - and the entire dataset: The table illustrates the number of matches, the market prediction accuracy (percentage of matches the outcome with lowest odds is the true outcome), the mean odds of the true outcome, the 1^{st} and 3^{rd} quartile odds and the mean odds for the three possible outcomes; home win, away win and draw. The calculations are based on *all* football matches in the English Premier League, French Ligue 1, Spanish LaLiga, German Bundesliga 1 and Italian Serie A.

	The complete dataset: 2017/18 - 2024/25	**Pre-Covid: 2017/18 until year-end 2019**	**Covid: year-beginning 2020 until year-end 2021**	**Post-Covid: year-beginning 2022 until 2024/25**
Number of matches	14,308	6,249	3,557	4,502
Market Prediction Accuracy	54.14%	54.54%	53.11%	54.40%
Mean true-outcome odds	2.97	2.96	3.00	2.94
1^{st} quartile odds	2.36	2.36	2.36	2.34
3^{rd} quartile odds	4.50	4.50	4.50	4.40
Mean home odds	3.02	3.03	3.13	2.91
Mean away odds	5.18	5.59	4.81	4.90
Mean draw odds	4.34	4.42	4.29	4.27

general public does not have. If such information would enable a strategic bettor to generate long-term positive returns in the betting market, it is an indication of the market not being of strong form efficiency, see section 2.4.1.

5.2 Covid-19: A change in regime

In many ways, the COVID-19 pandemic was not only a health crisis but also a profound societal and economic turning point. The effects went far beyond hospitals and lockdown policies, reshaping how humans go to work, spend time on leisure, socialize and consume entertainment. Among these effects; a seemingly significant shift in some main metrics in the betting markets.

In section 2.4.3, we explored the possibilities of constructing simple favorite and long-shot betting strategies, and we saw that the frequency of *very low* odds (large implied probabilities) was smaller in the post-Covid period, than in the pre-Covid period. Specifically, before 2020, there were approx. 11.22 bets with odds less than or equal to 1.09 per calendar-year, while on average there were 1.56 bets after year-end 2021 and 1.50 during the Covid pandemic (year-beginning 2020 until year-end 2021). While we have not found significant and persistent change in the long-shot strategy, we will investigate how the odds have changed throughout the period.

In table 5.1, main identification metrics of our dataset is illustrated based on four different periods: i.) the complete dataset, ii.) the pre-Covid period, iii.) during the Covid pandemic and iv.) the post-Covid period.

In Europe, football fans are passionate and will never shy away from yelling encouraging or, to put it lightly, discouraging things to the players on the field, however; suddenly the Covid-19 pandemic hit Europe in early 2020, and tens of thousands of loud spectators were substituted with quietness and perhaps the occasional bird. In table 5.1, the change in mean odds for home victories are the most noticeable. Mean odds for the home victories were significantly higher during the pandemic than in the pre- and post-pandemic period. Additionally, in the post-Covid period, the mean odds for a home victory have decreased 0.22 odds-points - to a level 0.10 odds-points lower than before Covid. Assuming the betting market is *somewhat* efficient, this indicates that the market have found an even stronger dependency on which team is at home and the result of the match after the pandemic.

Logically, as odds for a specific outcome decreases (home victory), the odds for the opposite outcome must increase (draw and away victory), however, this has not been the case. It seems that the market found itself with, roughly, the same prediction accuracy as before the pandemic, but with lower true-outcome odds and with lower odds for all 1X2 outcomes of matches. In other words, the betting market appears to have experienced inflation with odds on 1X2 outcomes decreasing across the board. This implies higher implied probabilities and thus *less* favorable pricing for back bettors. This is also prominent when investigating the 1^{st} and 3^{rd} quartile odds.

Recall, a bet exchange is a platform that sole purpose is to facilitate peer-to-peer betting, i.e. matching market participant that provide odds (lay) and market participants that buy odds (back). In this project, we have taken on the role as back bettors, while the change in market odds in table 5.1 indicate that it would have been interesting to take on the role as lay bettors.

This structural change in the provided odds on bet exchanges can be interpreted as a consequence of a fundamental shift in market sentiment, where odds have changed due to a change in underlying beliefs about uncertainty and general market confidence. After all, if no lay bettors provide bets, there is no bets to be wagered, and evidently, back bettors are still willing to engage in bets, even with lower odds.

For now, we assume we have some time before experiencing another global and all-encompassing pandemic. Therefore, we treat the pandemic as a period of outliers, and we argue that it will affect the modeling negatively if we include it in the training data. Therefore, we have excluded the 2020-21 season completely.

Another explanation as to why the odds have generally become lower on the betting exchange is the size of the bid-ask spread. It is not improbable that more and more people have started using the Betfair betting exchange in recent years, which should narrow the bid-ask spread, as more people provide more liquidity. This *should* not lead to lower odds in itself, unless there is a skewed orderbook, where the backers are more aggressive than the layers. Once again, we see this as a very probable possibility, as the backing orders are the equivalent to betting at traditional bookmakers, which is what most bettors are used to. Without detailed orderbook data from Betfair, this hypothesis is not possible to test, so for now it stays a hypothesis.

As our strategy is a backing-only strategy, we will actually add to this 'problem' of lower odds, because we will put additional pressure on the backing-side of the orderbook. We do not take this potential slippage into account in our backtest, as i.) we would not expect our market impact to be very high, and ii.) we would need much more detailed and granular data to actually have a chance to model our market impact.

5.3 Strategy results

Before going to the results from our machine learning based trading strategies, we will first outline the most important aspects of the ML model we ended up going with.

Obviously, we are using the XGBoost model, as this is the algorithm our homemade loss function was written to work with. Speaking of the homemade loss function, we choose a value of 1 for the hyperparameter c, which controls the amount of relative focus on decorrelation to the market for the model - we will talk more about c in section 5.4. We landed on this value, as we found it to consistently provide the best risk-adjusted returns.

We will not focus on the other hyperparameters of our XGBoost model, but for the interested reader, they are very briefly outlined in section A.2 of the appendix.

As for the features of the model, we have chosen the following seven features:

Feature	**Description**
Competition dummy	Indicates the competition
Mean points per match	
– This season	Mean points per match throughout the season
– Last 5 matches	Mean points per match in the last five matches
ELO rating	
– Absolute	Indicates momentary ability
– Rolling slope of the change	Indicates result momentum in the previous month
Days since last game	Indicates level of fatigue among players
Relegation distance	Points above relegation

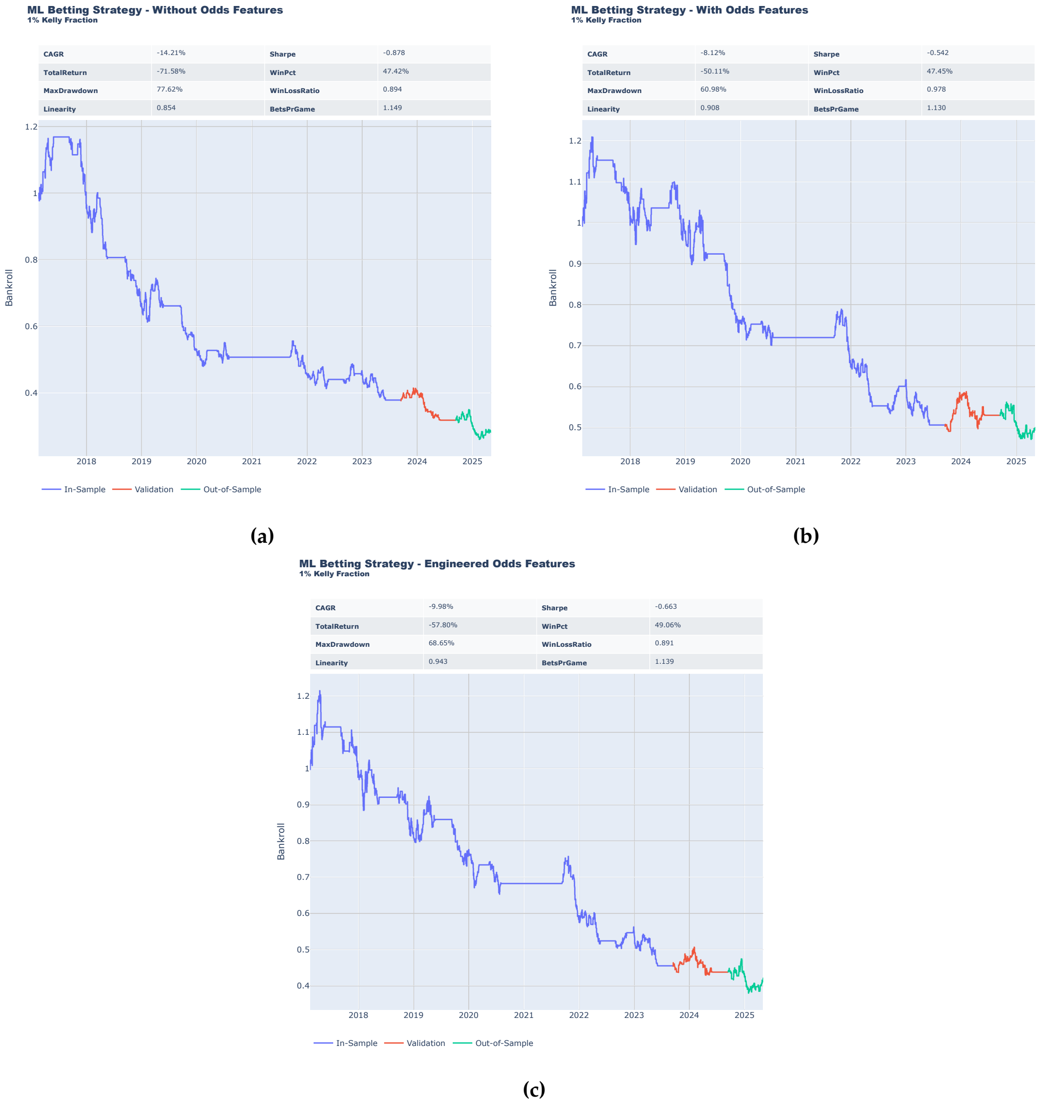
Logically, all the included features can be argued to be conceptually linked to the outcome variable; and while the numeric performance is *the* most important thing, it never hurts for the features to make sense intuitively.

In addition to the above features, we have gone back and forth on whether and how to include the odds from the betting exchange. Hubáček et al. (2019) also touch upon this dilemma, in that they find that including the odds improves the model's accuracy, while simultaneously increasing the correlation to the bookmakers probability estimate. In the end, these contrasting effects gave an inconclusive effect on profit generation. To test whether the betting exchange odds should be included in our model, we look at the results from 3 similar models, only differing in how we incorporate the odds from the betting exchange.

Before going through the results of the strategies, there are a few important things to note. First of all, even though the plots in the following section say 'In-Sample', none of the datapoints are really in-sample per-se. The results are calculated based on the rolling walk-forward cross-validation, as presented in section 4.3.1. By 'In-Sample', in this context, we mean that we have this piece of the dataset available for iterative testing, which means it could suffer from data-snooping. Additionally, it is worth mentioning that the metrics at the top of the figures are calculated based on *all* three periods.

The strategies presented in figure 5.1 all share the same basic features and hyperparameters presented above. Figure 5.1a does not incorporate any odds feature, figure 5.1b incorporates the home, draw and away odds directly, while figure 5.1c incorporates engineered odds features[1]. The strategies are made using the Fractional Kelly-Criterion as presented in section 3.2.1 - the maximum fraction of the current bankroll to bet is set to 1% on each outcome in order to decrease the strate-

[1] We have derived two features from the odds from the betting exchange and other bookmakers. These are described in section 3.1.3



CAGR	-14.21%	Sharpe	-0.878
TotalReturn	-71.58%	WinPct	47.42%
MaxDrawdown	77.62%	WinLossRatio	0.894
Linearity	0.854	BetsPrGame	1.149

(a)

CAGR	-8.12%	Sharpe	-0.542
TotalReturn	-50.11%	WinPct	47.45%
MaxDrawdown	60.98%	WinLossRatio	0.978
Linearity	0.908	BetsPrGame	1.130

(b)

CAGR	-9.98%	Sharpe	-0.663
TotalReturn	-57.80%	WinPct	49.06%
MaxDrawdown	68.65%	WinLossRatio	0.891
Linearity	0.943	BetsPrGame	1.139

(c)

Figure 5.1: Equity curves and strategy metrics for 3 betting strategies, utilizing odds features in different ways. All 3 strategies uses the same underlying XGBoost-model, only differing in how odds features are included. All 3 strategies bets according to the Kelly-Criterion, with a maximum Kelly Stake of 1% of the current bankroll.

gies' volatility and risk-of-ruin.

The first thing to notice for all three strategies in figure 5.1 is that the results are not exactly looking good. All strategies have a win percentage below 50% and a win/loss ratio below 1, which will always yield negative results.

While still generating highly negative returns, the two strategies which incorporate information from the odds clearly outperform the strategy which does not. Additionally, it would seem that the strategy in figure 5.1b which directly incorporates the odds from the betting exchange outperforms the strategy with the engineered odds features in figure 5.1c. This means that we, contrary to Hubáček et al., are able to conclude that including odds features in the model improves profit generation. The most probable reason for this conclusion is that we do not have as strong basic features as Hubáček et al. (2019), meaning that our accuracy simply takes too much of a hit if we do not somehow incorporate the information from the highly descriptive odds features.

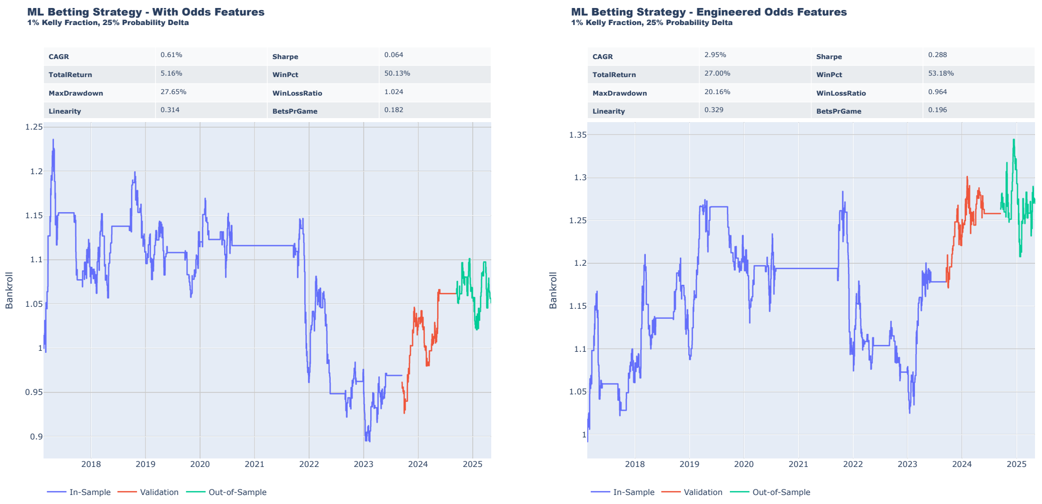
Of course, we cannot be satisfied with the results in figure 5.1; losing 8% of the bankroll every year is not exactly going to make anyone rich.

We do this by altering how we use the Fractional Kelly-Criterion slightly. The reason to use the Kelly-Criterion to size our bets, is that it is mathematically proven to maximize our long-term log of wealth. As mentioned in section 3.2.1, this does come with a few assumptions, though, the most significant one being that we know the true probability of an event occurring. This, of course, is not the case, which is why we came up with the idea to have a 'statistical confidence buffer'. What we mean by this, is that we do not immediately bet whenever our expected value is above 0; we only bet whenever our predicted probability of an outcome occurring is larger than ditto of the betting exchange plus some tuneable amount. Thus, the fraction of our bankroll that we bet on each outcome becomes:

$$k_i^f = \begin{cases} c \times \dfrac{e_i - 1}{o_i - 1}, & \textit{if } \hat{p} \geq \Delta \frac{1}{o_i}, \\ 0, & \textit{otherwise.} \end{cases}$$

, where k_i^f is the amount of the bankroll wagered on each outcome, c is the maximum Kelly Fraction, e_i is the *decimal edge* between the calculated probability of winning the bet and the market implied probability, o_i is the provided decimal odds of the outcome, $\hat{p}$ is our probability estimate of an outcome occurring, o_i is the decimal odds from the betting exchange and Δ is our 'statistical confidence buffer' (called the probability delta from here).

While this method is not grounded in a rigorous mathematical proof, it seemingly gives much more stable strategies, giving much improved metrics, and even a solid, positive annual growth rate, as can be seen in figure 5.2. For the strategies in figure 5.2 we found a probability delta of 25% to yield the best results. Apart from this additional probability delta, nothing has changed from the strategies in figure 5.1, which means that we are still using a 1% Fractional Kelly-Criterion to create our bets.



ML Betting Strategy - With Odds Features
1% Kelly Fraction, 25% Probability Delta

CAGR	0.61%	Sharpe	0.064
TotalReturn	5.16%	WinPct	50.13%
MaxDrawdown	27.65%	WinLossRatio	1.024
Linearity	0.314	BetsPrGame	0.182

ML Betting Strategy - Engineered Odds Features
1% Kelly Fraction, 25% Probability Delta

CAGR	2.95%	Sharpe	0.288
TotalReturn	27.00%	WinPct	53.18%
MaxDrawdown	20.16%	WinLossRatio	0.964
Linearity	0.329	BetsPrGame	0.196

(a)

(b)

Figure 5.2: Equity curves and strategy metrics two betting strategies - one using raw betting exchange odds directly as features and one using engineered odds features. Both strategies uses the same underlying XGBoost-model, only differing in how odds features are included. Both strategies bets according to the Kelly-Criterion, with a maximum Kelly Stake of 1% of the current bankroll and requires a minimum probability delta to that implied by the betting exchange of 25 percentage points.

What is quite clear this time, is that the strategy with the engineered odds features in figure 5.2b outperforms the strategy in figure 5.2a with raw odds features. Even though the strategy with the raw odds features has a win/loss ratio above 1, the 3 percentage points higher win percentage of the strategy with the engineered odds features gives a much higher overall annual growth rate of almost 3%. Apart from earning much more money, all other metrics are also improved - they bet approximately just as often, but Sharpe ratio and max drawdown is much improved.

For this reason, we will only be including the strategy with the engineered odds features for further analysis.

5.3.1 Statistical performance analysis

While especially the strategy in in figure 5.2b shows very promising signs, it makes sense to do a few tests to check the significance of this strategy.

In (Lopez de Prado, 2018, ch. 12) a lot of focus is put on the fact that the *actual* historical sequence of observations is purely a single scenario of what *could have* happened. An obvious test of whether our strategy is susceptible to a different order of its trades, is by computing confidence intervals with bootstrapping.

We boostrap by collecting the returns on all dates where we have made a bet. We then randomly sample from this pool of returns with replacement, until we have as many trading days as in the original strategy. We then do this 10,000 times, giving us a non-parametric distribution of our returns-series.

In figure 5.3, the bootstrapped strategy is displayed with a 90% confidence interval as well as the minimum and maximum bankroll for each trading day.

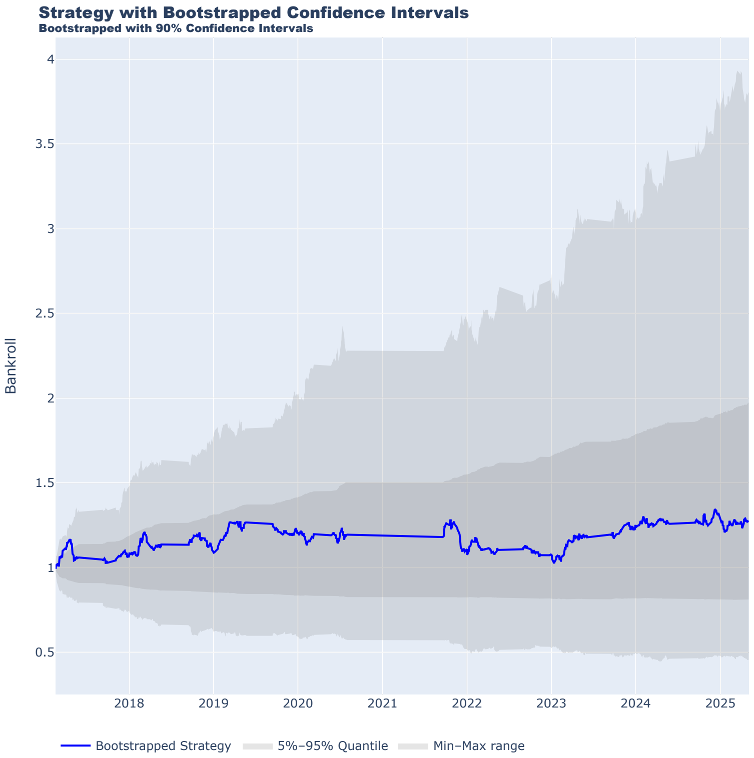


Figure 5.3: Equity curve of the same strategy as in figure 5.2b with boostrapped confidence intervals. The boostrap is done my iteratively sampling a fraction of 100% from the original strategy with replacement, 10,000 times. 90% confidence intervals are then computed as the 5th and 95th quantile for each trading day. The minimum and maximum values for each trading day from the boostrap is also plotted.

From the figure, it is quite obvious that the upwards potential outweighs the downside risk significantly. The 95% upper bound ends at a bankroll of 1.96, while the lower 5% bound ends in 0.81. Even though upwards potential dominates downside risk, the lower confidence interval is still below 0. This means that we cannot reject the null hypothesis that the strategy will end up losing money after 8 years (which is our backtesting horizon here). Further analysis of the bootstrapped distribution shows that the strategy is in negative territory after the 8 years in 20% of cases.

Heavy right tails

When looking at what kind of bets the strategy is the most dependent on winning, we see two opposing trends. The first one is that the strategy is quite dependent on winning quite few bets with quite large payoffs. A way to measure this, is by removing the 1% of bets with the largest payoff from the strategy. Here, we would like to see that the strategy is still profitable without these events, as these bets are on outcomes with a very low probability of happening. This means that these are the outcomes to which the underlying model - or we as strategy developers - are most prone to overfitting: rare events with a low likelihood of recurring in the future.

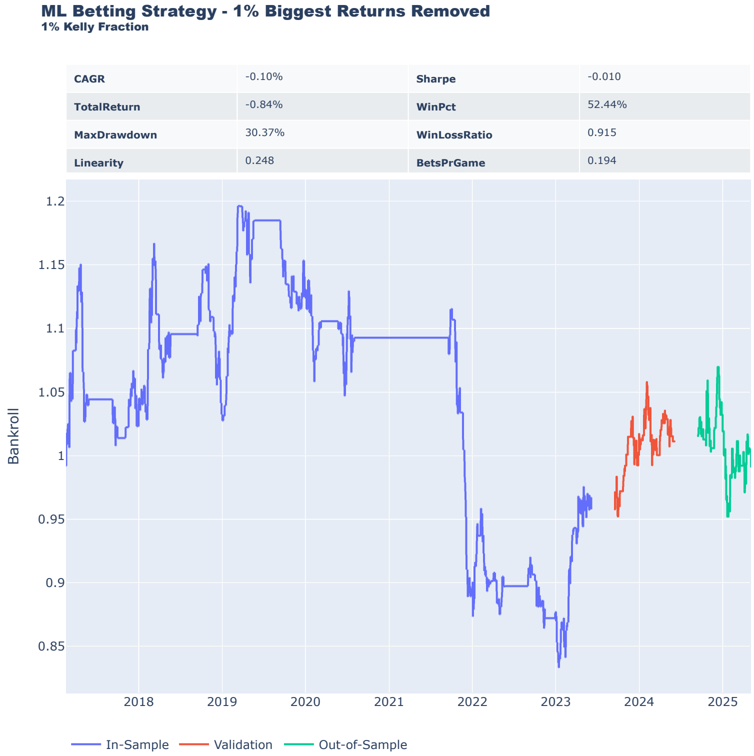
Figure 5.4 shows that removing the 1% biggest winners makes the strategy lose money in the backtesting period.

The counter point to this is that the bets we remove in figure 5.4 has an average odds of 1.83, which means that they are actually not very big outliers. Further analysis shows that the initial strategy actually wins most of it's money in a low odds range below 2, which is visualized in figure 5.5a. The strategy improves quite dramatically this way - mainly because the win/loss ratio stays mostly the same as in figure 5.2, but the win percent shoots up by 5.5 percentage points to 63%[2]

These findings tie nicely to the theoretical proposition of a *long-shot bias* as presented in section 2.4.3, which would mean that the odds on big underdogs are generally too low, compared to their actual probability of happening.

What makes this even more interesting is that the Betfair exchange seems to be providing the *lowest* odds compared to other bookmakers in a low odds range below 2, while it provides some of or even the highest odds in higher odds ranges, as can be seen in figure 5.5b. In the eyes of the authors, this fact makes the per-

[2] Given the relatively poor performance in the Out-of-Sample period, it can be discussed whether the strategy is overfit. We will tackle this problem in section 5.5



CAGR	-0.10%	Sharpe	-0.010
TotalReturn	-0.84%	WinPct	52.44%
MaxDrawdown	30.37%	WinLossRatio	0.915
Linearity	0.248	BetsPrGame	0.194

Figure 5.4: Equity curve and strategy metrics for the same betting strategy as in figure 5.2b, but with any returns bigger than or equal to the 99th quantile of returns removed.

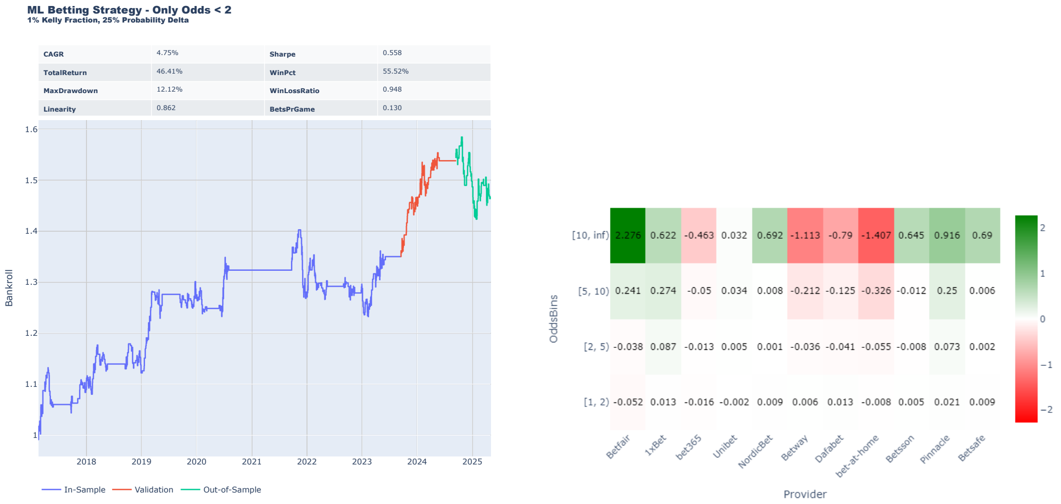
formance of the strategy *more* plausible, as our strategy could have earned significantly more money, if we based our profit on odds from different providers.

The one downside to the ′odds < 2′-strategy in figure 5.5a is that it does not earn money in the out-of-sample period, which could be a worrying sign that we have overfitted the model to the training data - we will return to this point in section 5.5.

5.3.2 Modern Portfolio Theory for strategy construction

Although the strategy utilizing the Kelly-Criterion performs relatively well, below we construct our portfolio of bets using the MPT procedure as outlined in section 3.2.2. The underlying ML-model used for constructing the probability predictions for the strategy is the same as in figure 5.1c - the ML model with the engineered odds features.

Additionally, we have chosen the γ risk-aversion hyperparameter from the MPT maximization problem in equation 3.4 to be 1. We found this value to give a good



CAGR	4.75%	Sharpe	0.558
TotalReturn	46.41%	WinPct	55.52%
MaxDrawdown	12.12%	WinLossRatio	0.948
Linearity	0.862	BetsPrGame	0.130

	Betfair	1xBet	bet365	Unibet	NordicBet	Betway	Dafabet	bet-at-home	Betsson	Pinnacle	Betsafe
[10, inf)	2.276	0.622	-0.463	0.032	0.692	-1.113	-0.79	-1.407	0.645	0.916	0.69
[5, 10)	0.241	0.274	-0.05	0.034	0.008	-0.212	-0.125	-0.326	-0.012	0.25	0.006
[2, 5)	-0.038	0.087	-0.013	0.005	0.001	-0.036	-0.041	-0.055	-0.008	0.073	0.002
[1, 2)	-0.052	0.013	-0.016	-0.002	0.009	0.006	0.013	-0.008	0.005	0.021	0.009

(a) Equity curve and strategy metrics for the same betting strategy as in figure 5.2b, but only including wagers on outcomes with odds below 2.

(b) Heatmap of the difference from the average odds across all providers for different odds providers in different odds ranges. A positive number means, that the average odds for that provider in that odds range is higher than the average for all providers in the odds range. The odds ranges are calculated based on the odds from Betfair. The Betfair odds have been multiplied by 0.95 to make them comparable to the odds from the traditional bookmakers.

Figure 5.5

balance between risk and return in the backtest.

The sizes of the bets in the strategies presented in figure 5.6 have all been rescaled ex-post in order for the average amount wagered to match their equivalent strategy using the Kelly-Criterion. This is done so we can compare the two strategies. While this introduces look-ahead bias, as we tweak the wagered amounts 'after' the bet has happened, the upside of being able to easily compare strategies across procedures easily outweighs the downside.

The reason why this introduces look-ahead bias is that we only calculate the average size once, at the end of the backtest, and use this to scale the historical size backwards. If we wanted to proceed without any look-ahead bias, we should have calculated the historical average size before each bet, and then used that average to scale the size of the *future* position. The problem with this implementation would be that we would not be sure to have perfectly comparable results to the Kelly-strategies, which is why we chose to go with the more simple implementation, where we only scale the size of the positions once, at the end.

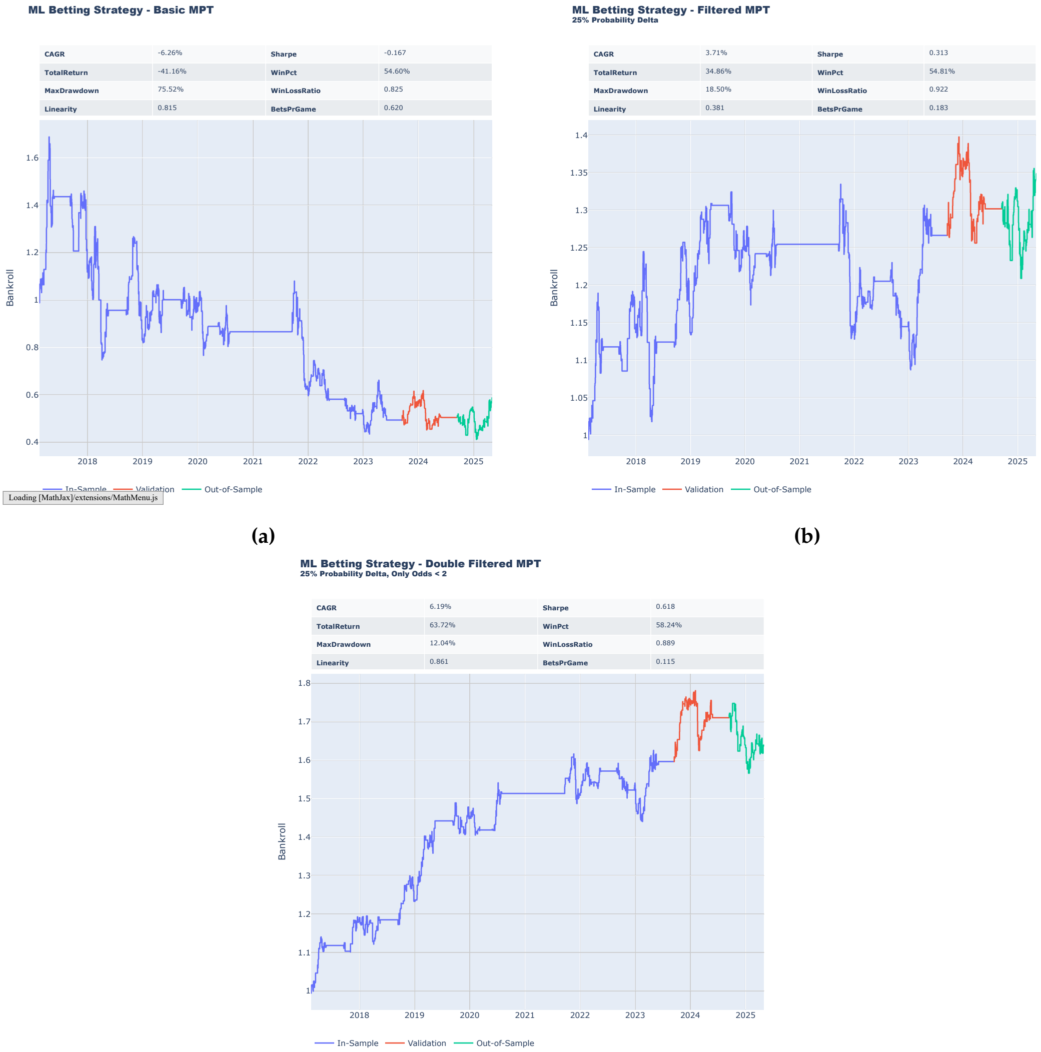
The strategy in figure 5.6a is based on the strategy in figure 5.1c and thus it does not have any additional filters after having constructed the positions with MPT. The strategy in figure 5.6b is based on the strategy in figure 5.2b, which means that the positions are constructed with MPT as before, but any bet where our probability estimate is not 25 percentage points higher than the implied probability of the betting exchange is filtered away. Lastly, the strategy in figure 5.6c is based on the strategy in figure 5.5a, which means that we impose an additional filter on the MPT-positions; in addition to the 25% probability delta, we now only bet on outcomes with odds < 2.

The strategy in figure 5.6a is the equivalent to the Kelly-based strategy in figure 5.1c, which means that it is the most basic strategy possible. What is the most interesting about this strategy compared to the one in figure 5.1c is that this only takes about half the amount of positions per game. Out of the box, the MPT strategy seems to be much better at 'choosing' the right outcomes to bet on, which is also clear from the 54.6% win percent of the strategy - an improvement of 5.5 percentage points compared to its Kelly-equivalent.

Still, the strategy is obviously not positive, as we lose about 6% of the bankroll every year. To rectify this, we move on to figure 5.6b - the Kelly-equivalent to this strategy can be found in figure 5.2b. What we do here is to impose a probability delta of 25% from the implied probability of the betting exchange before we are willing to take the bet.

The improvement from the simple implementation of MPT in figure 5.6a to figure 5.6b may not seem drastic, as the win percent only increases by 0.2 percentage points and the win/loss ratio improves by about 0.1. Still, as was the case for the Kelly-based strategy this improvement is enough to change the sign of the total return from negative to positive. If we compare to the equivalent Kelly-strategy in figure 5.2b, we once again see that the MPT-strategy performs better on all metrics apart from the win/loss ratio. It is especially impressive that the MPT-strategy ends with a higher total return and simultaneously has a smaller max drawdown.

Moving on to figure 5.6c, which has its Kelly-equivalent in figure 5.5a. This strategy is identical to the one in figure 5.6b, but this time we only allow the strategy to bet on odds smaller than 2. Like we saw with its Kelly-equivalent, this strategy shows by far the most promise, at least before the out-of-sample period where we once again see a pretty significant drop-off in performance. Compared to the Kelly-equivalent, this strategy has a much higher win percent of 58.24% compared to the 55.52% of the Kelly-based strategy.



ML Betting Strategy - Basic MPT

CAGR	-6.26%	Sharpe	-0.167
TotalReturn	-41.16%	WinPct	54.60%
MaxDrawdown	75.52%	WinLossRatio	0.825
Linearity	0.815	BetsPrGame	0.620

(a)

ML Betting Strategy - Filtered MPT
25% Probability Delta

CAGR	3.71%	Sharpe	0.313
TotalReturn	34.86%	WinPct	54.81%
MaxDrawdown	18.50%	WinLossRatio	0.922
Linearity	0.381	BetsPrGame	0.183

(b)

ML Betting Strategy - Double Filtered MPT
25% Probability Delta, Only Odds < 2

CAGR	6.19%	Sharpe	0.618
TotalReturn	63.72%	WinPct	58.24%
MaxDrawdown	12.04%	WinLossRatio	0.889
Linearity	0.861	BetsPrGame	0.115

(c)

Figure 5.6: Equity curves and strategy metrics for betting strategies based on the modern portfolio theory procedure as presented in section 3.2.2. The three strategies in the figure is based on the same procedures as the strategies in figure 5.1c, 5.2b and 5.5a, respectively.

All in all, it seems that the MPT-based strategy consistently achieves a higher win percentage than the Kelly-based strategy, which similarly achieves a higer win/loss ratio. The MPT strategies does seem to provide a higher return with a similar or smaller max drawdown, which suggests that the risk-adjusted returns are superior for these strategies.

5.4 Does the decorrelation loss function actually work?

The trading strategies presented in this chapter revolves around a machine learning model with the loss function presented in section 4.2.

To recap, we are using the below loss function to build our XGBoost model:

$$-\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\hat{p}_i) + (1 - y_i) \cdot \log(1 - \hat{p}_i) + c \cdot (\hat{p}_i - 1/o_i)^2\right] \tag{5.1}$$

, where the hyperparameter c is a penalty parameter, i.e. the value determining the algorithm's relative focus on decorrelation compared to minimizing the log-loss.

As presented previously, we ended up using a value of 1 for c, as that is what we found to consistently provide the best balance between returns and risk, i.e. the risk-adjusted return. As the above loss function is relatively central to the thesis, we also want to provide a brief analysis of the effect of c.

Hubáček et al. (2019) use a Monte Carlo simulation to prove that a higher decorrelation between a bettor's probability estimates and the implied probability estimates of the bet exchange yields a higher return. What they do not test, though, is whether a higher c actually *does* give a higher decorrelation to the market in the first place.

To test this, we set up a small experiment using real data instead of a simulation. We use the exact same model as has been used the entire chapter, only changing two things: The value of c and as we have done previously, which predictors we use; no odds features, raw odds features or engineered odds features. To calculate profit for the models, we use the procedure from figure 5.2b, where we have a probability delta of 25% and a fractional Kelly-approach, wagering a maximum of 1% of the current bankroll on a single outcome.

The results in table 5.2 shows very interesting results, which both support and contradict our initial hypothesis. The results presented in table 5.2 is computed for the entire backtesting period, also including the validation and out-of-sample periods. If we look at the CAGR, it is once again obvious that the model needs some kind of odds features to have positive returns. Additionally, we see that for the model with the raw odds features and the engineered odds features, the returns are increasing until $c = 1$. This is exactly as we would expect, as the effect from additional decorrelation outweighs the loss from worse accuracy, until a certain point.

Features	**c**	**CAGR**	**Accuracy**	**Odds Corr.**	**Odds Dist.**
Without Odds	0.00	-0.017	0.505	0.812	0.111
	0.50	-0.108	0.522	0.881	0.150
	1.00	-0.129	0.523	0.879	0.153
	1.50	-0.073	0.520	0.861	0.162
	1.75	-0.221	0.501	0.697	0.297
With Odds	0.00	-0.019	0.519	0.875	0.095
	0.50	0.012	0.534	0.935	0.140
	1.00	0.006	0.533	0.935	0.141
	1.50	-0.001	0.531	0.928	0.143
	1.75	-0.148	0.521	0.805	0.260
Engineered Odds	0.00	-0.016	0.512	0.863	0.099
	0.50	-0.040	0.530	0.926	0.142
	1.00	0.029	0.531	0.924	0.143
	1.50	-0.035	0.528	0.913	0.148
	1.75	-0.125	0.515	0.788	0.267

Table 5.2: Table showing the effect of the decorrelation parameter c when paired with different sets of features. The presented results is for the entire backtesting period, including the validation and -out-of-sample periods.

This does not line up with the results for the accuracy and odds correlation of the models - in these columns we see the exact opposite of what we would expect. As c grows, accuracy *increases*, while odds correlation *decreases* - at least until c reaches a value of 1.5, where we actually see the expected results.

Why do we see this? An explanation for the accuracy could simply be training data overfitting. If the ML model is overfitting to the training data, it would make sense that less focus on actually minimzing the log-loss term of the loss function could end up giving a higher accuracy.

As for the increasing odds correlation as c increases, the explanation could be even more simple. Even though Hubáček et al. (2019) (and we) call c the 'odds correlation' hyperparameter, it is actually *not* what we are working with. If we look at the c term in equation 5.1, it is actually the squared difference between the probability estimate of the model and the implied probability of the market, which we are attempting to drive up by increasing c. For this reason, the 'Odds distance' column has been added to table 5.2, which shows this measure - and here the results are more intuitive. We see that as c increases, the squared difference between the model's probability estimate and the implied probability increases, which is perfectly in line with what the loss function seeks to do.

Therefore, this exercise shows us two things. 1) An (additional) sign that we may have overfitted our model to the training data, and 2) the c hyperparameter does actually *not* decorrelate the probability estimate with the market probability. Instead it increases the squared difference between the two estimates, which in the end gives the expected results; a higher c yields better results, until the penalization of similarity influence the accuracy too much, and the profit starts to drop off.

5.5 Did we overfit the strategies to the training data?

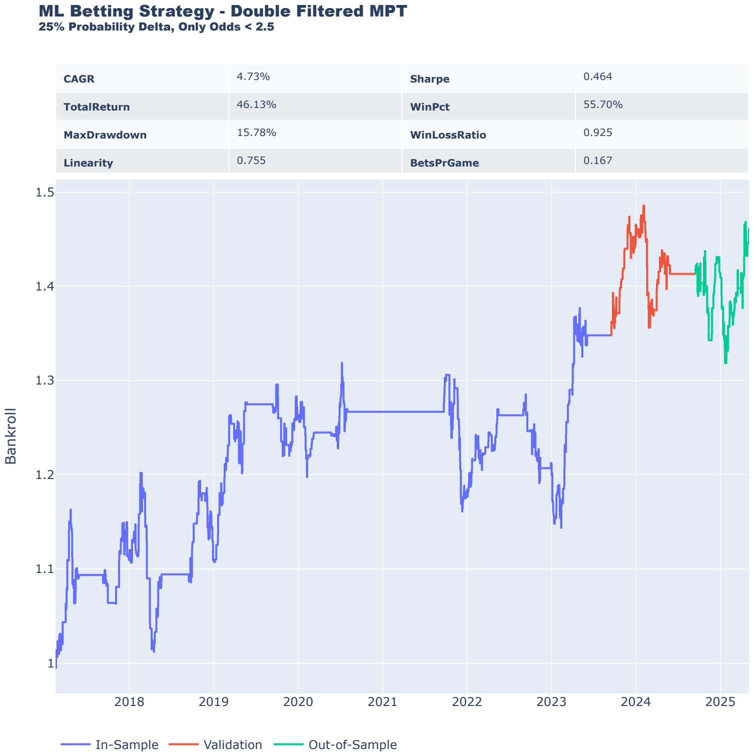
We have already mentioned a few times that there are a few signs that the betting strategies have been overfit to the training data. Even though it may be unwise to admit to overfitting, it is still a topic we feel is important to explore. As was written in section 4.3, backtest overfitting can be considered scientific fraud, which is why we feel it is our duty to investigate whether we have fallen into any traps.

The first thing that suggests overfitting was when working with the strategies in figure 5.2b and figure 5.6c, where we only bet on an outcome if our probability estimate is more than 25 percentage points larger than the implied probability of the market and if the odds is below 2. While this makes sense intuitively (we only bet when we disagree a lot with the market and we only bet on relative favorites (section 2.4.3)), the strategies mimics an almost perfectly linear curve right until the out-of-sample period - this is quite incriminating evidence.

There can be a few possible explanations. It could of course purely be bad luck, and the strategy could bounce back in the future, but the much more plausible explanation is overfitting. This does not necessarily mean that the entire underlying strategy should go in the trash, though. A good check to make is whether a small change in a trading- or hyperparameter would change the conclusions of the strategy significantly.

In figure 5.7 we use the MPT strategy from figure 5.6c, but this time we only want to bet when the odds are below 2.5 instead of 2.

The yearly growth rate changes quite significantly from more than 6% to just under 5%, but still the strategy seems to be quite solid - and now it even makes money in the out-of-sample period. A (very fair) counter point here is of course that we are



ML Betting Strategy - Double Filtered MPT

25% Probability Delta, Only Odds < 2.5

CAGR	4.73%	Sharpe	0.464
TotalReturn	46.13%	WinPct	55.70%
MaxDrawdown	15.78%	WinLossRatio	0.925
Linearity	0.755	BetsPrGame	0.167

Figure 5.7: Equity curve and strategy metrics for the same strategy as in figure 5.6c, but this time the odds can be up to 2.5 instead of 2, in order for us to make a bet.

now choosing the best trading parameters for the out-of-sample set ex-post, which is 'not allowed'. Still, the strategy without this parameter altogether (the strategy where the only restriction is the probability delta), yields positive results, which could mean that the 'odds < 2'-rule could simply be a step too far, and a sign that the less ideal, but perhaps more obtainable results of figure 5.6b is what is actually realistic to achieve in real life.

The second case where we fear overfitting is when going through the c hyperparameter, where a larger c gave higher accuracy for the model, which does not really make intuitive sense, unless the amount of information in the data is lackluster. While it might be the case that our ML algorithm is slightly overfitted, that does not mean that a trading strategy *based* on its predictions necessarily is completely overfitted. We will argue that overfitting is a spectrum, and even though the underlying ML model may be overfit to some degree, that does not mean that it cannot produce signals which can be useful for a trading strategy to be built on top of it.

5.6 Football: a high-variance sport

The Danish philosopher Søren Kierkegaard (1813-1855) argued that passion is fundamental for a meaningful existence. Across the world, football embodies passion, history and community - an identity and feeling of belonging that mirrors Kierkegaard's on profound engagement in life. This passion resonates regardless of skin color, social class or political beliefs, making it a widespread and influential element of culture in modern Europe.

However, it would perhaps have been wiser to put passion and interest to the side, and focused on another sport. Football is fundamentally a high-variance sport; i.) goals are few, ii.) on-field punishments are detrimental for the course of the match and iii.) the stakes are high in *almost* every match throughout the season.

i.) In the dataset of this project, on average 2.80 goals are scored per match. Compared to various other *ball*-based sports[3], the number of goals scored in a match of football is much lower, i.e. it is arguable that each individual football goal has more significance to the result of the match compared to other sports.

ii.) Without engaging with the regulatory framework of various sports, football is unique as it is *normal* to have a player removed from the match if given a red card. In this case of a red card, the team of the player that receives the red card has one player less on the field in the rest of that match. In handball and ice-hockey, players can be sent off and not be substituted by another player immediately. Specifically, in handball, the suspension lasts 2 minutes and in ice-hockey the suspension lasts either 2, 5 or 10 minutes - but in both sports, the penalized team will ultimately regain full strength.

iii.) In most football leagues globally (the exceptions are the American, Candian, Australian and Indian leagues), the worst ranked team(s) are relegated to a lower-level division. As a consequence, the worst ranked teams have the utmost motivation to avoid losing matches, and therefore, the vast majority of all matches are of importance throughout the season.

In a match between the highest and lowest ranked team in the league, it can be assumed that both sides are highly motivated to fight for the league title and survival, respectively. While the favored team is assigned the highest implied probability of

[3] American Football, basketball, baseball, ice-hockey, handball, tennis, table tennis, volleyball, cricket, etc.

victory, a non-negligible chance of the underdog scoring one or even two goals persists in all football matches. Suddenly, the likelihood of winning the match is significantly reduced for the favored team. Additionally, suppose a player on the favored team is shown a red card in the beginning of the match; the favored team must then play the rest of the match with one less player on the field. This is also likely to fundamentally change the dynamics of the match.

In essence; all football matches are subject to stochastic outcomes in each match that can critically change the match and the implied probabilities of outcomes. These are tendencies that are not experienced to the same degree in various other sports. In other sports, a single goal will seldom significantly influence the implied probabilities of outcomes and teams will only experience a different number of players on the field momentarily. Perhaps the match might not even matter for one of the teams as relegation does not exist and the title is mathematically out of reach.

As implicitly mentioned throughout the project, the betting market is believed to have *all* (or at least almost all) information implemented into the odds for football matches, and therefore, the above-mentioned structural intricacies of football is implemented into the odds. Events of *very* low odds are inherently rare in competitive football matches, as uncertainty and potential stochastic events that change the match dynamics remain more probable than in other sports, where the greater amount of goals might make the result more likely to converge to the expected outcome.

5.7 Subconclusion on results

In this chapter we have presented a lot of different strategies with mixed results and analyzed and discussed their results. The final thing we want to discuss, is *which* strategy we would like to pick, if we were to use it going forward.

As mentioned in section 5.3.2, we generally feel that the MPT strategies perform better than their Kelly-based equivalents, so this method would be our starting point. In addition, we feel the development potential for MPT is more vast than for a Kelly-based strategy, which is more 'set in stone'. For MPT, we can tweak the way we put together games to form a portfolio, how much weight we put on risk vs. return and we can even begin playing with the maximization problem and it's constraints. We do not see the same potential for the Kelly-Criterion, which seems much less flexible.

In terms of the actual strategy we would choose in the future, the choice is between the MPT strategy in figure 5.6b (probability delta of 25%) and the MPT strategy in figure 5.6c (probability delta of 25% and only odds < 2). As mentioned in section 5.5, the strategy in figure 5.6c perhaps took it a 'step too far' in terms of hard-coded trading signals, as the performance deteriorates in the out-of-sample period. For this simple reason, we are most inclined to choose the more simple strategy with only the probability delta filter in figure 5.6b, even though its overall performance is not quite as good.

In the end, this begs the question - does it even make sense to run a strategy with an annual growth rate of 3.71% per year? A yearly return of under 4% on the stock markets would not exactly be deemed a world-class return, so why would it make sense to put our hard-earned money toward this betting strategy, instead of putting them in an index fund?

For us, there are two answers to the question. Firstly, covariance and diversification; as a rule of thumb the covariance between the return on our betting strategy and, for example, the return on the stock markets is zero. As presented in the equations section 3.2.2, this low covariance will give a lower variance to an overall portfolio of both stocks and the betting strategy.

Secondly, the betting strategy can be used as a *tool* to get out as much money from so-called 'freebets', which we mentioned very briefly in chapter 2. A freebet is where a bookmaker will give you a free sum of money (to lure you onto their platform) - in turn you will often have to bet on a minimum number of games before being able to withdraw the money. Even if we deem the betting strategy to have too weak results to bet with systematically, it would still be very useful to maximize the payout from a freebet.

6 Conclusion

This thesis set out to explore whether it is possible to construct a profitable betting strategy on the Betfair betting exchange for the 5 biggest European football leagues by leveraging advanced machine learning methods.

To do this, we extend on the methods from Hubáček et al. (2019), as we implement a XGBoost model with a classification-specific loss function aimed at decorrelating our prediction with the market-prediction - the amount of decorrelation is controlled by the hyperparameter c. We implement this loss function, as we would simply lose the commission of the betting exchange, if our forecast is identical to that of the market. We implement this model to predict the probability of each outcome of each game occurring.

In order to construct betting strategies with these probability estimates, we use the very classic (Fractional) Kelly-Criterion, which helps us find outcomes with positive expected values based on our probability estimates and the market odds, and size our bets based on these. Additionally, we once again expand on the methods of Hubáček et al. (2019), as we propose an implementation of Modern Portfolio Theory (MPT) for betting, where the covariance between the outcomes of each game is incorporated.

Empirically, we find that the decorrelation-based machine learning technique works well, even though it technically does not work quite as expected. We find that a relatively large c of 1 works best, as it provides the highest returns. What we also find, though, is that a higher c surprisingly did *not* lead to more decorrelation. Rather, it lead to a higher squared difference between our prediction and the implied market prediction - which is also exactly what the loss function is intended to do. During this exercise we also conclude that it is beneficial to include the odds from the market in some way, even though it made our model more correlated with the market prediction. In the end, we found that features *derived* from the raw odds features performed the best, as these strike a nice balance between high amounts of information, without increasing our correlation to the market too much.

When translating the probability estimates into strategies, it was clear that both the Kelly-Criterion and MPT strategies could not stand on their own, as a new naive implementation of these strategies lose 6%-15% of the bankroll each year. For this reason, we implement some simple rule-based filters to enhance performance.

The first rule that we implement is that our probability estimate must be 25% higher than the implied probability of the market before placing a bet. While this rule filters away most of the drawdowns and makes total return positive, we cannot reject the null hypothesis that the strategy will yield negative result using bootstrapped confidence intervals.

For this reason, we implement an additional rule based on our initial exploration of the betting market and the human psychology behind it. Here we find that bettors on average have a 'long-shot bias', which means that they favor betting on outcomes with a low probability and high payoff. This should in theory drive down the odds on low-probability outcomes, which in turn should drive up the odds on high-probability outcomes.

We sought to exploit this effect by limiting our strategy to only bet on outcomes with odds < 2. Both in-sample and in our validation set, this strategy performs exceptionally well, but unfortunately the strategy loses a significant amount of money in the out-of-sample period, giving us a worrying sign that the strategy might be overfit to the training data.

In the end, we conclude that the MPT-based strategies are the most favorable, as they provide a just as high return as the Kelly-based strategies, but with lower risk and shallower drawdowns. Additionally, we conclude that the strategy with both of our rule-based filters, which significantly underperformed in the out-of-sample period, has too high a risk of being overfit. For this reason, we choose the MPT-based strategy with a rule of only placing bets when our probability estimate is more than 25% bigger than that of the market as our best strategy.

We argue that even though this strategy 'only' has an annual growth rate 3.71%, which is not exactly high compared to investing in stocks, it would still make sense to wager money on it in real life. First of all, the strategy is in theory completely decorrelated with the return on other assets, which would make it a way to diversify a portfolio.

Additionally, the strategy can be used to maximize the payoff from the freebets often offered by bookmakers. This would make the strategy highly valuable, even if we were to decide that it is not robust enough for systematic, real-world deployment.

Bibliography

[Brandt et al. 2009] Brandt, Michael W. ; Santa-Clara, Pedro ; Valkanov, Rossen: Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns. In: *The Review of Financial Studies* 22 (2009), Nr. 9, 3411–3447. `http://dx.doi.org/10.1093/rfs/hhp003`. – DOI 10.1093/rfs/hhp003

[Brooks 2019] Brooks, Chris: *Introductory Econometrics for Finance, FOURTH EDITION*. Cambridge University Press, 2019. – 1–696 S. `http://dx.doi.org/10.1017/9781108524872`. `http://dx.doi.org/10.1017/9781108524872`. – ISBN 9781108524872

[Bruce et al. 2012] Bruce, A. C. ; Johnson, J. E. ; Peirson, J.: Recreational versus professional bettors: Performance differences and efficiency implications. In: *Economics Letters* 114 (2012), 2, Nr. 2, S. 172–174. `http://dx.doi.org/10.1016/j.econlet.2011.10.014`. – DOI 10.1016/j.econlet.2011.10.014. – ISSN 01651765

[Buchdahl 2003] Buchdahl, Joseph.: *Fixed odds sports betting : statistical forecasting and risk management*. High Stakes, 2003. – 224 S. – ISBN 1843440199

[Chen 2022] Chen, Tangqi: *XGBoost Parameters — xgboost 2.0.3 documentation*. `https://xgboost.readthedocs.io/en/stable/parameter.html`. Version: 2022

[Chen & Guestrin 2016] Chen, Tianqi ; Guestrin, Carlos: XGBoost: A Scalable Tree Boosting System. Version: 2016. `https://github.com/dmlc/xgboost`. 2016. – Forschungsbericht

[Cortis & Levesley 2016] Cortis, Dominic ; Levesley, Jeremy: Betting Markets: Defining odds restrictions, exploring market inefficiencies and measuring bookmaker solvency. 2016. – Forschungsbericht

[Deschamps 2007] Deschamps, Bruno: BETTING MARKETS EFFICIENCY: EVIDENCE FROM EUROPEAN FOOTBALL. 2007. – Forschungsbericht

[Direr 2012] Direr, Alexis: Are Betting Markets Efficient? Evidence from European Football Championships. (2012), 9. http://dx.doi.org/10.1080/00036846.2011.602010{ï}. – DOI 10.1080/00036846.2011.602010ï

[Dixon & Pope 2004] Dixon, Mark J. ; Pope, Peter F.: The value of statistical forecasts in the UK association football betting market. In: *International Journal of Forecasting* 20 (2004), 10, Nr. 4, S. 697–711. http://dx.doi.org/10.1016/j.ijforecast.2003.12.007. – DOI 10.1016/j.ijforecast.2003.12.007. – ISSN 01692070

[Düring et al. 2022] Düring, Bertram ; Fischer, Michael ; Wolfram, Marie T.: An Elo-type rating model for players and teams of variable strength. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380 (2022), Nr. 2224. http://dx.doi.org/10.1098/rsta.2021.0155. – DOI 10.1098/rsta.2021.0155. – ISSN 1364503X

[Fama 1970] Fama, Eugene F.: Efficient Capital Markets: A Review of Theory and Empirical Work. 1970 (2). – Forschungsbericht. – 383–417 S. – ISBN 130.225.53.20

[Franck et al. 2010] Franck, Egon ; Verbeek, Erwin ; Nüesch, Stephan: Prediction accuracy of different market structures - bookmakers versus a betting exchange. In: *International Journal of Forecasting* 26 (2010), 7, Nr. 3, S. 448–459. http://dx.doi.org/10.1016/j.ijforecast.2010.01.004. – DOI 10.1016/j.ijforecast.2010.01.004. – ISSN 01692070

[Harari 2011] Harari, Yuval N.: *Sapiens: A Brief History of Humankind*. 2011

[Hegarty & Whelan 2024] Hegarty, Tadgh ; Whelan, Karl: Estimating Expected Loss Rates in Betting Markets: Theory and Evidence. Version: 2024. https://bookies.com/guides/what-is-the-vigorish. 2024. – Forschungsbericht

[Hillier et al. 2012] Hillier, David. ; Grinblatt, Mark. ; Titman, Sheridan.: *Financial markets and corporate strategy*. McGraw-Hill Higher Education, 2012. – 854 S. – ISBN 9780077129422

[Hubáček & Šír 2023] Hubáček, Ondřej ; Šír, Gustav: Beating the market with a bad predictive model. In: *International Journal of Forecasting* 39 (2023), 4, Nr. 2, S. 691–719. http://dx.doi.org/10.1016/J.IJFORECAST.2022.02.001. – DOI 10.1016/J.IJFORECAST.2022.02.001. – ISSN 0169–2070

[Hubáček et al. 2019] Hubáček, Ondřej ; Šourek, Gustav ; Železný, Filip: Exploiting sports-betting market using machine learning. In: *International Journal* *of Forecasting* 35 (2019), 4, Nr. 2, S. 783–796. http://dx.doi.org/10.1016/J.IJFORECAST.2019.01.001. – DOI 10.1016/J.IJFORECAST.2019.01.001. – ISSN 0169–2070

[Jabin et al. 2015] Jabin, Pierre-Emmanuel ; Junca, Stéphane ; Junca, Stéphane A.: A Continuous Model For Ratings. In: *Continuous Model For Ratings. SIAM Journal on Applied Mathematics* 75 (2015), Nr. 2. http://dx.doi.org/10.1137/140969324{ï}. – DOI 10.1137/140969324ï

[James et al. 2023] James, Gareth ; Witten, Daniela ; Hastie, Trevor ; Tibshirani, Robert: An Introduction to Statistical Learning with Applications in R Second Edition. 2023. – Forschungsbericht

[Kahneman & Tversky 1979] Kahneman, Daniel ; Tversky, Amos: Prospect Theory: An Analysis of Decision under Risk. Version: 1979. https://about.jstor.org/terms. 1979 (2). – Forschungsbericht. – 263–292 S. – ISBN 130.225.53.20

[López 2022] López, Colin: WHAT'S THE LINE? THE INFLUENCE OF NUMERICAL LITERACY ON THE PERCEPTIONS AND EVALUATIONS OF SPORT ODDS. 2022. – Forschungsbericht

[Malkiel 1973] Malkiel, Burton G.: A RANDOM WALK DOWN WALL STREET The Time-Tested Strategy for Successful Investing BURTON G. MALKIEL. 1973. – Forschungsbericht

[Mcmillan 2018] Mcmillan, David G.: PREDICTING STOCK RETURNS Implications for Asset Pricing. 2018. – Forschungsbericht

[Lopez de Prado 2018] Prado, Marcos M. d.: *Advances in Financial Machine Learning*. First. Wiley, 2018. – ISBN 978–1119482086

[Rathke 2017] Rathke, Alex: An examination of expected goals and shot efficiency in soccer. In: *Journal of Human Sport and Exercise* 12 (2017), Nr. Proc2. http://dx.doi.org/10.14198/jhse.2017.12.proc2.05. – DOI 10.14198/jhse.2017.12.proc2.05. – ISSN 1988–5202

[Shang et al. 2021] Shang, Xuesong ; Duan, Hebing ; Lu, Jingyi: Gambling versus investment: Lay theory and loss aversion. In: *Journal of Economic Psychology* 84 (2021), 6. http://dx.doi.org/10.1016/j.joep.2021.102367. – DOI 10.1016/j.joep.2021.102367. – ISSN 01674870

[Snowberg & Wolfers 2010] Snowberg, Erik ; Wolfers, Justin: *Explaining the Favorite-Longshot Bias: Is it Risk-Love or Misperceptions?* 4 2010

[Sun 2020] Sun, Dennis: *Introduction to Probability*. `https://dlsun.github.io/probability/`. Version: 8 2020

[Wilkinson & Klaes 2017] Wilkinson, Nick ; Klaes, Matthias: An Introduction to Behavioral Economics, Nick Wilkinson, Matthias Klaes (2017). (2017)

[Winkelmann et al. 2024] Winkelmann, David ; Ötting, Marius ; Deutscher, Christian ; Makarewicz, Tomasz: Are Betting Markets Inefficient? Evidence From Simulations and Real Data. In: *Journal of Sports Economics* 25 (2024), 1, Nr. 1, S. 54–97. `http://dx.doi.org/10.1177/15270025231204997`. – DOI 10.1177/15270025231204997. – ISSN 15527794

[Wooldridge 2012] Wooldridge, Jeffrey M.: *Introductory Econometrics: A Modern Approach*. 5th. 2012

[xgboost developers 2022] xgboost developers: *Demo for creating customized multi-class objective function*. `https://xgboost.readthedocs.io/en/stable/python/examples/custom_softmax.html`. Version: 2022

A Appendix

A.1 Derivation of the probability of the product of two non-independent discrete random variables

The derivation of the probability of the product of two non-independent discrete random variables P_i and P_j:

$$E[P_iP_j] = -b_ib_j\left[\hat{p}_i(o_i-1)+\hat{p}_j(o_j-1)\right]+\hat{p}_kb_kb_j \tag{3.10}$$

, where P are the profits from betting on any two outcomes i and j of the three outcomes in the 1X2 market of a match, o are the decimal odds, b are the wagered amounts and $\hat{p}$ are the estimated probabilities of the specific outcome happening. k is the third outcome.

The derivation is based on Sun (2020, ch. 25), which utilizes 2D LOTUS (2-dimensional Law of the Unconscious Statistician) for two discrete random variables. The derivation has its foundation in the function:

$$g(P_i, P_j) = P_i \times P_j \tag{A.1}$$

, which is the product of the profits/losses P_i and P_j, and the joint mass probability function f, which is the joint probability of the outcomes for P_i and P_j occurring. A table gives the best intuitive understanding:

$P_j \backslash P_i$	$b_i \times -1$	$b_i(o_i-1)$
$b_j \times -1$	$\hat{p}_k$	$\hat{p}_i$
$b_j(o_j-1)$	$\hat{p}_j$	0

The table shows the probability of different payoffs for betting on either outcome j or i.

To give an intuitive example, say that outcome i is betting on a home win and outcome j is betting on an away win. The probability of losing both of these bets (payoff of $b_i \times -1$, the entire stake times the wagered amount) is $\hat{p}_k$, the probability of a draw. The probability of winning both bets (payoff of $b(o-1)$) is 0, as both the home team and away team cannot win at the same time. The probability of winning one of the bets and losing the other is $\hat{p}_i$ and $\hat{p}_j$, respectively.

Therefore, when combining the two functions, we get:

$$E[P_iP_j] = \sum_{P_i}\sum_{P_j} g(P_i, P_j) f(P_i, P_j) = \sum_{P_i}\sum_{P_j} P_iP_j f(P_i, P_j) \tag{A.2}$$

, which is an aggregated combination of equation A.1 and the joint mass probability function f. The estimated probabilities and payoffs for betting on the 1X2 market in matrix A.1 can be computed based on equation A.2. Suppose we only focus on one 1X2 bet:

$$\begin{aligned} E[P_iP_j] &= (-b_i \times -b_j) \times \hat{p}_k + ((b_i(o_i - 1)) \times -b_j) \times \hat{p}_i + (-b_i \times (b_j(o_j - 1))) \times \hat{p}_j \\ &= b_ib_j\hat{p}_k + (-b_ib_j(o_i - 1) \times \hat{p}_i) + (-b_ib_j(o_j - 1) \times \hat{p}_j) \\ &= b_ib_j\hat{p}_k - b_ib_j\hat{p}_i(o_i - 1) - b_ib_j(o_j - 1) \\ &= -b_ib_j[\hat{p}_i(o_i - 1) + \hat{p}_j(o_j - 1)] + \hat{p}_kb_kb_j \end{aligned} \tag{A.3}$$

A.2 Hyperparameters of the XGBoost model

Hyperparameter	**Value**	**Role**
c	1	Relative weight of decorrelation term
colsample_bytree	0.9	Fraction of features sampled for each new tree
gamma	0	Min. loss reduction required for a further split of a tree
learning_rate	0.1	Amount of shrinkage applied to each update
max_depth	3	Maximum depth of each tree
n_estimators	500	Number of trees
subsample	0.6	Random fraction of training set sampled for each boosting round

Table A.1: XGBoost hyperparameters, their values and a brief explanation of their role. For further reading on the XGBoost hyperparameters, refer to Chen (2022)