DEPARTMENT OF COMPUTER SCIENCE

Stylized Facts in Online Sports Betting Exchanges

[redacted]

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Engineering in the Faculty of Engineering.

Thursday 4th May, 2023

Abstract

I analyse UK horse racing markets time series data from Betfair - an online sports betting exchange, and derive a range of stylized facts. I identify the unconditional distribution of log returns as a generalised Gaussian. Unlike the distribution of financial asset returns which is known to exhibit heavy tails, the distribution of returns from a betting exchange shows signs of light tails with a tail index closer to that of a Gaussian than a power law. Statistical testing and analysis show a lack of gain-loss asymmetry in this distribution. The absolute returns are concluded to be difference stationary while log returns stationary. Linear autocorrelations appear to be negligible with exception to for the first few lags. Nonlinear autocorrelations show faster decay than that typical of financial time series, indicating a lesser degree of volatility clustering. Lastly, the estimation of the Hurst exponent affirms the absence of long-range memory in betting exchange time series, which points to evidence of mean reversion. Overall, the results suggest higher informational efficiency than that typically exhibited by financial markets, according to the stylized facts about them.

Contents

List of Figures

List of Tables

Ethics Statement

This project did not require ethical review, as determined by my supervisor, Dave Cliff.

Chapter 1

Introduction

In economics and finance a stylized fact is a simplified presentation of an empirical finding or some statistical characteristics. While stylized facts capture the broad tendencies in the data, they might no longer hold true at a more detailed glance. As said by Robert Solow there *"is no doubt that they are stylized, though it is possible to question whether they are facts"*[81]. However, when properly derived, they remain useful for motivating the construction of economic models and validating them.

Since their emergence in the early 2000s online betting exchanges have transformed the gambling industry. Platforms like Betfair, Smarkets of Betdaq attract millions of bettors [88]. Instead of a traditional bookie, it is bettors that both back and lay their bets on a market. The data structure underpinning betting exchanges is very similar to a Limit Order Book, which aggregates and anonymises the orders received. This makes online sports betting exchanges very similar to financial exchanges.

1.1 Motivation

Stylized facts in financial markets have long been a topic of research and a number of statistical properties have been well established as true for most financial time series. Some of those facts include heavy tails in the unconditional distribution of returns, volatility clustering, slow decay of nonlinear autocorrelation and the so-called leverage effect. Those findings are a result of over half a century of robust research. Online sports betting exchanges have not even existed for the majority of that time and even during the last 20 years have not been studied with the objective of deriving stylized facts nearly as extensively. It is this gap in knowledge, paired with the prevalence and popularity of online sports betting exchanges, as well as their similarity to financial markets that motivates this project. New information on betting exchanges could thus provide novel insights into financial markets.

Another opportunity that arises from identifying the stylized facts of sports betting data is facilitating the development of more realistic synthetic data generators. In economics, one of the most important indicators of the validity and correctness of a model is whether it explains or reproduces the stylized facts of the subject. The same would be true for synthetic data generators: in order for the produced data to be deemed realistic it ought to exhibit the same statistical characteristics as the real-world data it is replicating. An example of such a generator would be the Bristol Betting Exchange[20], which simulates in-play horse-racing betting markets, as well as the races themselves. Synthetic generators of betting exchange data are invaluable to the development and analysis of mathematical or algorithmic approaches to profitable betting or trading on these exchanges [5, 6, 13]. Algorithmic trading strategies utilizing machine learning need large amounts of data for training and validation. Currently data from real sports betting exchanges can be purchased in relatively small batches, which is expensive and inconvenient. Furthermore, some Machine Learning approaches may need more data than there is available at all. Finally, aside from its purely academic potential, the project could be of great interest to any betting algorithm users, as it facilitates their development and improvement. Algorithmic bettors are used by both betting companies and regular bettors, just like robot traders populate most major financial markets.

1.2 Aims and objectives

The main objective of the project is to rigorously analyse high-frequency data from a real-world online sports betting exchange and identify its statistical characteristics. The properties that are of interest in this project are especially the ones present in financial time series, as the differences between sports betting exchanges and financial markets would be of great interest. Therefore I aim to:

1. Analyse the distribution of returns
2. Identify time dependence properties of returns. Those include:
   (a) Linear and nonlinear autocorrelation
   (b) Volatility clustering
   (c) Long-range memory
   (d) Leverage effect
3. Provide the derived stylized facts on horse racing markets to be reproduced by the Bristol Betting Exchange
4. Draw conclusions on the nature of online sports betting exchanges

Chapter 2

Background

2.1 Online Sports Betting Exchanges

Betting exchanges operate by offering a marketplace for bettors to bet on an outcome of an event. Punters can both back and lay an outcome of an event, meaning they can either bet on or against the outcome occurring. Users can offer and request odds from other bettors and the betting exchange platform operates as a matching engine. Additionally, betting exchanges permit wagering on derivative bets as well in a practice known as trading. Derivative bets allow trading on the movement of the odds during an event, rather than simply betting on the outcome.

One of the defining innovations of betting exchanges is in-play betting. In-play betting allows the punters to place bets not only before the start of a sporting event, but also throughout its duration, often until its very end. The odds for each market are updated in real-time, reflecting the current state of the event. This addition was popularised by one of the industry pioneers, Betfair.

2.1.1 The Basics of Betting: Odds, Overround and Betting On an Online Exchange

An online betting exchange offers the opportunity of either backing or laying each of the N possible outcomes of an event. The price on an outcome is known as odds and is represented by the fixed ratio between the stake and the returns on a winning bet. The odds are usually displayed in a fractional or decimal format. This is best illustrated by an example: if the odds are 4:1 in fractional or 5.0 in decimal format, the backer can win 4 times the amount they wagered known as the stake, plus they get their stake back as well. If the outcome does not happen the backer loses their stake.

The inverse of the decimal odds on an outcome can be interpreted as the probability of that outcome. The sum total of all the odds on the outcome of a single event is known as the overround. Typically those proposed probabilities sum to more than one, which allows for making a profit. Customers interact with the information offered by the exchange via either a grid or a ladder interface. Each row in a grid is dedicated to one runner - a participant of the event, and each cell contains specific odds and the total volume wagered on them. While the grid view is more easily readable for an inexperienced user, the ladder view shows the full market depth, meaning the complete array of odds and the liquidity at each price. The ladder interface allows for quicker reactions to changes in the market and thus is favoured by traders operating in volatile markets.

2.1.2 History

Online betting exchanges emerged in the early 2000s following the wide expansion of Internet access into homes and workplaces in the early 1990s, as well as the development of first gambling software by companies like Playtech and Microgaming, and the development of encrypted communication protocols allowing for secure monetary transactions by the likes of CryptoLogic and Netscape in the mid-1990s[90]. The first online sports betting exchanges, Betfair and Flutter.com, both launched in the year 2000 and merged a year later. While Flutter.com was the better-funded first mover, Betfair managed to quickly dominate the market and end up with a market share of nearly 90%[16]. Betfair remains the largest of the existing online betting exchanges and as of march of 2023 attracts over 20M monthly visitors[79].

Online sports betting in general is a rapidly growing industry with an expected compound annual growth rate (CAGR) of more than 12.0%. Betting exchanges make up around 20%[39] of the revenue share of the sector. This number does not fully reflect the importance of betting exchanges, as their profit margins are considerably lower than that of traditional bookmakers, more on which in the next section.

Lopez-Gonzalez, H. Griffiths, M.D.[61] list the standardization, homogenisation and numerical quantification of sport performance, the globalization of sporting markets, the increasing mediatisation of sport contents, as well as the increasing ubiquity of personal communication technologies as the main cultural propagators of online betting exchanges' growth in prominence.

2.1.3 Betting Exchanges versus Traditional Bookmakers

Betting exchanges are a disruptive alternative to the traditional bookmaker model of betting. In the latter, a bookmaker, i.e. a "bookie" sets the odds on an event and accepts bets from customers. Therefore bettors can only back bets as each bet is laid by the bookmaker. Unlike a betting exchange, the bookmaker takes a risk on each bet. This is compensated for by the generally less favourable odds offered, as well as taking a cut of the bettors' wins. In comparison, a betting exchange's sole source of profit is a commission on winning bets. As a result, betting exchanges offer significantly more competitive prices and hence higher returns than traditional bookmakers[22].

Yet despite offering better odds, betting exchanges have not eliminated traditional bookmakers. On the contrary, traditional bookmakers and online betting exchanges formed a mutually beneficial relationship[16]. The first reason for that is relatively straightforward: the two types of platforms attract different groups of customers. The number of bettors who prefer the simplicity and familiarity of traditional bookmakers is sufficiently large to sustain the model. Secondly, betting exchanges are more informationally efficient and lead price discovery[22]. This allows bookmakers to hedge their positions at a lower cost by drawing insight from betting exchanges. Finally, betting exchanges proved to attract sophisticated high rollers - gamblers who consistently bet large sums of money and successfully employ advanced trading strategies and hedging algorithms, who happen to be the least profitable of the bookmakers' customers. Drawing them away only helps traditional bookmakers.

Bookmakers include firms, such as William Hill or Ladbrokes, but they can also be individuals (licensed dealers).

2.2 Similarities Between Betting Exchanges and Other Market Types

2.2.1 Financial Markets

Betting exchanges bear a lot of similarities to financial markets. The very mechanism of betting on an online sports betting exchange is directly comparable to trading on a stock market in a sense that both bring together a number of traders willing to buy some asset in a particular amount and a number of traders looking to sell said asset on a centralised platform. What is more, both betting exchanges and financial markets provide a similarly structured view of the current state of the market, which includes information on the supply and demand of the market.

Furthermore, the data structure underpinning a betting exchange, known as a market for a particular event, is a close relative of the Limit Order Book (LOB), which lies at the heart of a financial market. A Limit Order Book is a record of outstanding buy and sell orders, as well as the prices and quantities of the assets. The traders can submit and cancel their orders. The sellers issue orders to sell a certain number of units of their assets at a certain price (the asking-price) and the buyers do the same with orders to buy (bid). The exchange's matching engine will then show the asks as a list of tuples comprised of the of price and quantity of the asset being sold, ordered by price best to worst and aggregate the orders which share the same price. The bid side of the Limit Order Book is then analogous, only displaying ordered buy orders instead. The matching engine will match the corresponding buy and sell orders and the unmatched orders will be cleared. A betting exchange's market matches bettors who back and lay by odds and stake in an analogous fashion .

2.2.2 Prediction Markets

Another type of markets that online sports betting exchanges resemble are prediction markets. A prediction market's purpose is to generate a forecast about an outcome of a future event based on the beliefs of its participants. PredictIt and Smarkets are both examples of online prediction markets. It is most common for prediction markets to use binary options: traders on the market buy and sell contracts at market price $p \in [0, 1]$. If the event occurs, the buyer receives one from the seller, otherwise they receive zero.

While sports betting exchanges have been used by researchers to predict the outcomes of sporting events[70, 77], forecasting event outcomes is not their main focus and they are rather inefficient at achieving this goal compared to actual prediction markets[50].

2.3 Stylized Facts in Financial Time Series Data

As previously established, betting exchanges very closely resemble financial markets. Financial markets have been around for much longer and have attracted the attention of large numbers of researchers. The research on the microstructure of financial markets includes a wide body of work dedicated to identifying and analysing the stylized facts of financial time series data[68, 25, 21].

A few properties of financial time series data are worth noting before describing their stylized facts. Firstly, financial market data is by nature discrete - the smallest price change allowed is know as a tick. In the stock market, a tick is typically one cent. Additionally, the transactions on financial exchanges do not occur at equally spaced time intervals. Finally, multiple transactions can occur at the same time, as the units that the time at the exchange is measured in might not be small enough to capture the interval between some transactions, especially in periods of heavy trading.

2.3.1 Autocorrelation

Autocorrelation, or serial correlation in the case of discrete time, is a measure of correlation of a variable with a lagged version of itself over a certain time interval.

It has been established that in liquid markets price changes (and therefore log returns) exhibit an insignificant degree of autocorrelation. In that case, lack of autocorrelation is a sign of an efficient market[86]. In high frequency data some evidence of negative first-order autocorrelation of returns has been found, one that is slowly decaying over time, meaning slowly it gets less and less significant with the size of the lag[78, 58]. A number of possible explanations for that phenomenon have been proposed, the most popular being the bid-ask bounce[12]. The bid-ask bounce occurs when the price of an asset bounces between the ask price and the bid price. Another commonly listed reason is order imbalance[34]. Order imbalance is an imbalance between the number of buy and sell orders which leads to the inability of the market to match them. A negative autocorrelation means simply that a positive return in one period will be followed by a negative one in the next period and vice versa. A positive autocorrelation indicates the opposite. While the returns themselves appear not to show significant positive autocorrelation, the absolute returns or their squares do display a strong positive autocorrelation with slow decay that exhibits power-law behavior[85, 19]. The autocorrelation of absolute returns or their squares is also referred to as nonlinear autocorrelation.

Taylor Effect

The autocorrelation of returns $|r|^p$ is the highest when $p = 1$[27]. This observation follows the finding of a 1986 study by Taylor, S. J.[84] where he concludes that absolute returns $|r|$ exhibit higher autocorrelation than than the squared ones $|r|^2$, and hence was termed the Taylor Effect.

2.3.2 Volatility Clustering

Volatilty clustering is perhaps the most famous of the stylized facts of financial time series. It has been known since the 1950s[52]. The effect manifests itself in small changes of either sign being followed by small changes and large changes being followed by large changes. In other words, the market volatility tends to cluster in time. It indicates shock persistence and is ubiquitous in financial markets.

Volatily clustering is characterized by Autoregressive Conditional Heteroskedasticity (ARCH) - a model developed in 1982 by Engle[29], and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) developed by Bollerslev 4 years later[11]. Both models consider the current error variance as a function of the previous period's error variance and are quite successful in modelling historic time series data. They do struggle however with forecasting future changes, where an AutoRegressive Fractionally Integrated Moving Average (ARFIMA) process[3] proves to be more accurate.

2.3.3 Leverage Effect

Returns and future volatility have been shown numerous times to be negatively correlated[10]. This relationship is known as the leverage effect. To better model the leverage effect, modified versions of GARCH have been developed, such as the GJR-GARCH[37], NG-GARCH[76], and QR-GARCH[67]. The effect has its roots in the way financial leverage works. Negative returns increase financial leverage, which makes the asset riskier and therefore leads to a short-term increase in volatility[10]. The leverage effect is another measure of nonlinear dependence in returns and can be measured by the correlation of returns with subsequent squared returns.

2.3.4 Stationarity and Seasonal Effects

A time series is said to be stationary if its statistical properties remain constant over time and do not exhibit a trend. Strong stationarity requires the mean, variance and autocorrelation function of a time series to be constant. Second-order stationarity, or weak stationarity only restricts the requirement to the mean and auto-covariance functions. Financial time series prices have been found not to be strongly stationary due to the presence of trends, seasonalities and other effects such as volatility clustering. These effects make the standard deviation of returns vary over time. The returns are however assumed to be weakly stationary[45]. There is some evidence however, that those properties might not hold true under ceratain conditions. As en example, Lee at al.[57] find real stock price indices are stationary in developed and developing countries to be stationary.

Seasonality

Seasonal patterns, meaning patterns that repeat at fixed intervals of time, can be found in various aspects of financial time series including volatility, trade frequency, volume, and spreads. Their presence has long been documented, with Wachtel finding seasonal patterns in stock returns for the first time in 1942[87].

In financial markets the seasonality is especially strong in intra-day and intra-week data. The volatility during the day exhibits a U-shaped pattern[66], with the highest volatility soon after opening and right before the closing of the exchange. The reason for the early increase in volatility is the market participants analysing the market overnight and trading first thing in the day. The increased volume drives the volatility increase. Traders then wait until later in the day to observe the behaviour of the market and adjust their positions close to the end of the trading day. Similarly, the day-of-the-week effect, also present in most financial markets, is a phenomenon where the daily returns vary depending on the day, with the last day of the trading week observing the highest returns.

2.3.5 Unconditional Distributional Properties of Returns

The distribution of financial time series data has been long known not to follow a normal distribution[63]. The kurtosis K, which is a measure of the combined weight of a distribution's tails relative to the center of the distribution, of a Gaussian distribution is such that $K = 0$. For the unconditional distribution of returns from a financial market the kurtosis takes a positive value, indicating heavy tails. The tails remain heavy, albeit a little less, even in a conditional distribution, that is one corrected for volatility clustering.

The tail of a distribution can be described using the tail index. The tail index is defined as the negative of the exponent of the tail of the distribution. The tail index of a normal distribution is infinite, as the tails of a normal distribution show exponential decay. In contrast, the tail index in high frequency financial time series unconditional distribution of returns has been empirically shown to range from 2 to 5[23]. Therefore the tails can be described as exhibiting a power-law or Pareto decay.

Aggregational Gaussianity

Aggregational Gaussianity is the phenomenon of returns tending to normality as the time scale over which returns are calculated increases[28]. Patterns such as fat tails or skewness tend to cancel out or average out when returns are aggregated over longer time periods. However counterintuitive, aggregational Gaussianity has been widely accepted as a stylized fact of financial time series data. According to Bingham and Kiesel[9], in general, returns gathered over a term in excess of 16 days typically conform to normality.

Gain-Loss Asymmetry

Gain-loss asymmetry states that the positive price changes are more frequent, but negative returns are bigger, meaning the losses have a stronger impact on financial outcomes[72, 21]. It has been found to be most pronounced for short horizons[47]. The phenomenon has been linked to the concept of loss aversion which can be explained by prospect theory, where loss aversion refers to the tendency for people to prefer avoiding losses over acquiring gains. It as been suggested that traders can be twice as sensitive to losses as to gains[49].

Chapter 3

Related Work

Thanks to their rapid raise in popularity, online betting exchanges have attracted the attention of researchers from the fields of economics, finance, computer science, econophysics and more. Therefore since their emergence in the early 2000s online sports betting exchanges have been studied from a number of perspectives.

3.1 A Brief Review of Various Work on Online Sports Betting Exchanges

As an example, a significant amount of work has been dedicated to studying the behaviour of human bettors. A large proportion of research in this area appears to focus on problem gambling and its impact, including studies by Griffiths[41], Braverman and Shaffer[30], Gray et al. [40] and many more. Those studies often aim to not only capture the characteristics of problem gamblers' activity on online betting exchanges, but also the properties of the gambling platforms that influence and drive their behaviour. Examples include a 2022 paper by Hing et al.[44], where they qualitatively assess the impact changes in online betting platforms over the past ten years had on individual gamblers' habits. As mentioned by Cliff[20], there is also a large number of empirical studies on human bettors' behaviour utilising the concept of a representative bettor and often drawing from prospect theory. Those include works by Swidler and Shaw[83] and Suhonen et al.[82].

Other studies describe their results as stylized facts about human bettors' behaviour. In an experiment with regular sports bettors from 2022 Chegere et al.[18] have found that under the sports framing subjects are more likely to overestimate their chances of winning. The experiment was not conducted using an online betting exchange however, but fixed-odds bets with a bookmaker and focused on low income bettors only.

Another seemingly popular topic for research on betting markets is the use of mathematical or algorithmic approaches to profitable betting or trading on those exchanges. Examples include: Aruajo-Santos[5]; Bunker and Susnjak[13]; Axen and Cortis[6].

Finally, several works have been written on sports gambling platforms' role within the economy and integration with the digital, sporting and gambling sectors. Examples include Koning, R.H. and van Velzen conducting a SWOT-analysis of online betting exchanges[54] and Casadesus-Masanell, R. and Campbell, N.[15] analysing the relationship and competition between Betfair and traditional bookmakers. Another article by Lopez-Gonzalez, H. & Griffiths, M.D.explores the integration of social and technological processes that enabled online sports betting to proliferate, as well as the market integration of online betting with other neighbouring industries[61].

3.2 Stylized Facts of Online Sports Betting Data

All things considered, online sports betting exchanges are far from an uncharted territory in terms of academic research, yet to the best of my knowledge, there has been very little work done on the topic of stylized facts in online sports betting exchanges in the vein of the vast landscape of research on stylized facts in financial time series data, which was briefly outlined in section 2.3. The stylized facts of financial markets have been studied since as early as the 1960s, with Mandelbrot identifying heavy tails in the distribution of asset returns in 1963[63]. While online sports betting exchanges have only been around for a fraction of that time, the gap in knowledge is still glaring.

3.2.1 Favourite-Longshot Bias

That being said, one stylized fact of online sports betting exchanges has been well established and widely explored, namely the presence and nature of the favourite-lonsghot bias. The favourite-longshot bias is a long-studied phenomenon present in both financial and betting markets. It can be described as bettors overvaluing the longshots - outsiders, contestants that would be considered unlikely to win, and relatively underestimating the favourites - participants with the highest likelihood of winning. As expected, the body of work on this bias in sports betting exchanges is not as large as that for financial markets, nevertheless there are numerous examples, which include a 2022 paper by Koning and Zijm[55]; a 2020 thesis by Kan[51]; a 2016 study by Abinzano, Muga and Santamaria[2] and a paper on the favourite-longshot bias in UK football markets by Cain, Law and Peel[71]. Most of them agree the favourite-longshot bias is especially pronounced in sports betting data. Its causes however remain a topic of debate, with some scholars interpreting it as a sign of market inefficiency, while others attribute it to insider information some traders might potentially possess.

3.2.2 Informational Efficiency

Some research examines the informational efficiency of sports betting markets more broadly. Using football data from Ladbrokes, William Hill and Betfair, Croxson and Reade[22] find betting exchanges to be more informationally efficient than dealer markets, as well as on average offering significantly more competitive pricing. Similarly Franck, Verbeek and Nuesch[35] conclude betting exchanges offer superior prediction accuracy in comparison to other market structures. Matching results are reported by numerous other studies, including the aforementioned 2007 paper by Koning and Zijm[55].

Meier et al.[65] examine whether sports betting exchanges are semi-strong form efficient i.e., whether new information is incorporated into betting prices in a rapid and complete manner, and conclude them not to be, at least temporarily. The data used in the study came from a football market and evidence of a bias towards the home team's winning probability. Smith, Paton and Williams[80] explore the favourite-longshot bias, as well as more broadly market efficiency using data from UK horse racing markets. They conclude that unlike traditional bookmakers and financial markets, exchanges exhibit both weak and strong form market efficiency.

3.2.3 Other stylized Facts

The attempts at deriving the kind of stylized facts that would be comparable to the ones well established in financial market data are few and far between. One of them comes from Hardiman et al.[42]. Their work focuses on long-range correlations and limits its scope to football markets. They distinguish between in-play and half-time data and find evidence of long-range correlations in the magnitude of returns. Unlike in the case of financial time series, the trading volumes are concluded to be a short-memory process. Finally, a self-affine mean-reversion of the implied probability at half-time is revealed via Detrended Fluctuation Analysis (DFA).

The only attempt at deriving a truly wide range of stylized facts from sports betting data I have encountered comes from PA Bebbington. In his 2017 PhD thesis[7] he presents stylized statistical facts for in-play horse racing markets. Using the data from over 12736 races collected from Betfair he comes up with a number of observations. His results consisted of the moments of the price increment signals not being of a normal distribution and showing a highly leptokurtic structure, the dispersion in the odds increasing on average throughout the duration of an event, and gamblers failing to rank the horses accurately when their number was significant. He also concludes the market to be informationally efficient by comparing the empirical values of the implied odds and true winning probabilities. The favourite-longshot bias is confirmed to be present in sports betting markets once again. A somewhat surprising element of his analysis is the range of measures of the dispersion of odds employed in an attempt to explore the evolution of uncertainty. It includes the Gini index, Theil index and Generalised entropy index. Broadly speaking, the study appears to be approaching the research topic from the perspective of econophysics and his methods differ from the conventional economic time series analysis practices.

3.3 Stylized Facts of Prediction Market Data

As previously explained in section 2.2.2, prediction markets are the closest relatives to sports betting exchanges, perhaps closer than financial markets. Some of the statistical tendencies present in financial market data have been found not to hold true in online political prediction markets. Having analysed data from 3385 events, Restocchi[75] found that the distribution of returns in a prediction market is right skewed. In financial time series the opposite is usually true[72, 21]. Restocchi's 2018 thesis is however the only study on the topic that I am aware of. Furthermore, while similar in many regards, prediction markets and sports betting markets do differ[56]. A very important distinction concerns the binary nature of the assets traded on the examined market. The contracts traded on PredictIt are Arrow-Debreu securities, i.e., the bettor wins either 1 dollar or 0 depending on the outcome of the event. As thoroughly explained in the paper itself, this necessitates the use of raw returns in the analysis of the distribution and time dependence of the returns. Sports betting exchanges, like financial markets, do not typically feature binary options and thus are better suited to having their returns analysed using logarithmic or absolute returns.

Nevertheless, considering the range of properties and phenomena explored in the paper as well as the methods used I personally consider it to be the one work that is the most relevant to the aims and objectives of this thesis. Similarly to Restocchi, I aim to derive stylized facts for a type of exchange that is similar to financial markets, but has not been studied much from that angle before.

Chapter 4

Data

4.1 Initial Data Format

The data used for the analysis was purchased from Betfair. The betting platform offers detailed time-stamped data from the exchange. There are three differently priced package plans: PRO, BASIC and ADVANCED. They differ by both frequency and content. The data in this study comes from the PRO plan, which is the most comprehensive one with the widest scope of content provided at the highest, namely tick-by-tick frequency.

The Betfair Historical Data files are available for download as `tar.bz2` files. The directory contains the data from a month of trading and within the main folder there are directories separating the data by event. The event folders contain `.bz2` files, each corresponding to one market for that event.

The market files contain market data in json format. The actual content of the files is a list of market change messages documenting every change in the market. The two types of messages recorded are `MarketDefinition` and `RunnerChange` messages. The former define the details of the market and document any adjustments made to it, such as the number of active runners, whether it is currently in-play or its status. The latter describe changes to the details of a runner, namely prices. Every time any price on any of the runners changes, a new `RunnerChange` message is sent. The detailed description of the message format is shown in tables 4.1, 4.2 and 4.3.

Abbreviation	Description
op	Operation type
clk	Sequence token
pt	Published Time - in milliseconds since the start of the epoch
mc	The market change follows this field
id	The market's unique identifier

Table 4.1: The start of a market change message. These fields are included in every message and are followed by either a market definition (table 4.2) or a runner change message (table 4.3)

4.2 Preprocessing

While Betfair offers an API for accessing data purchased from the Betfair Historical data service as well as a pythonic wrapper written in C and Rust for using it[36], accessing the json files directly proved to be a more efficient and convenient option. The unpacked data from all the files was transformed into Pandas DataFrames, which were then filtered and processed to produce multiple CSV files containing the relevant information. Since the unpacked data totals to less than 600MB, saving and loading data was unlikely to cause much of a bottleneck in the execution time. Therefore I chose the familiarity and readability of the CSV format over some faster alternatives. Had the data analysis required larger amounts of data, one of the many available binary file formats like pickle or feather would have been a more appropriate choice.

The runner change messages from all the files were grouped into a single file as well as split into multiple separate files, one per each event for convenience. The market definition messages were stored separately. One market definition CSV file contains all market definition messages with the purpose of recording any changes in the markets' details throughout the events' duration, while another stores one

Abbreviation	Description
Id	The unique identifier of the market
Venue	The name of the venue the event is held at
bspMarket	Whether the market supports Betfair SP betting or not
turnInPlayEnabled	Whether the market is set to turn in-play or not
persistenceEnabled	Whether the market supports 'Keep' bets if the market is to be turned in-play
marketBaseRate	The commission rate applicable to the market
eventId	The unique identifier for the event
eventTypeId	The unique eventTypeId that the event belongs to
numberOfWinners	The number of winners in the market
bettingType	The betting type of the market. One of ODDS, ASIAN_HANDICAP_DOUBLE_LINE or ASIAN_HANDICAP_SINGLE_LINE
marketType	The market base type
marketTime	The market start time
suspendTime	The market suspend time
bspReconciled	Whether the market starting price has been reconciled
complete	Whether runners can still be added to the market or not
inPlay	Whether the market is currently in play
crossMatching	True if cross matching is enabled
runnersVoidable	True if runners in the market can be voided
numberOfActiveRunners	The number of currently active runners
betDelay	The number of seconds an order is held until it is submitted into the market. Orders are usually delayed when the market is in-play
status	Market's status, for example OPEN or SUSPENDED
regulators	Market's regulators
discountAllowed	Whether the users' discount rate is valid on the market or not
timezone	The timezone the event is taking place in
openDate	The start and end dates of the event. By default GMT
version	A number indicating market changes
name	Market's name
eventName	The name of the event

Table 4.2: Market Definition fields

Abbreviation	Description
status	The status of the selection, for example ACIVE or LOSER
sortPriority	Runner's sort priority
bsp	Runner's Betfair Starting Price
removalDate	Date and Time of runner's removal
id	The unique selctionId of the runner
name	Runner's name
hc	Runner's handicap
adjustmentFactor	The adjustment factor applied upon selection's removal
tv	The total amount matched across the market, the Traded Volume
ltp	Last Traded Price
spb	Starting Price Back
trd	Traded PriceVol
spf	Starting Price Far
atb	Available to Back
spl	Starting price Lay
spn	Starting price Near
atl	Available to Lay

Table 4.3: Runner Change fields

message per market for a more concise view. Another independent CSV file extracted from the market definition information records the winner as well as the number of runners for each market. Finally, the last set of CSV files produced from the original data contains raw time-stamped returns from all the markets and events.

4.2.1 Runner Changes

The runner change data is obtained by filtering for runner changes (`rc`) in all the market change messages. The format of the json files is such that everything following the `mc` field, be it a runner change or a market definition, gets loaded into the DataFrame as a single string, which needs to be parsed to extract the desired fields. I then convert the `pt` field to normal Unix epoch time in seconds and save it as a string of GMT time in the format of `Year-Month-Day T Hour:Minute:Second`. The `inPlay` field indicating whether the market is currently in-play is not present in runner change messages and has been added from the information contained in the market definition messages, based on their timestamps and market identifiers. The DataFrame is then sorted by time in ascending order. After dropping the fields I am not interested in, I am left with the final data format shown in table 4.4. It is worth noting that most of the fields in market change messages are delta based and nullable, meaning they will be set to null if they are not changed.

Field	Description	Data Type
atb	Available to Back	string
id	Selection identifier	int
t	Time	string
inPlay	Whether the market is currently in-play	bool
spn	Starting Price Near	float
spf	Starting Price Far	float
atl	Available to Lay	string
spl	Starting Price Lay	string
trd	Traded PriceVol	float
ltp	Last Traded Price	float
tv	Traded Volume	float
spb	Starting Price Back	string
eventId	Unique event identifier	int
marketId	Unique market identifier	float

Table 4.4: The final structure of the runner change data

4.2.2 Market Definitions

As previously mentioned, the market definition messages have been exported to two separate files: one containing all market definition changes, and one with only single market definition per market. Both files however have almost the same structure, which has been outlined in table 4.5. The only difference is the presence of an `inPlay` field in the full market change file. It is of type boolean and indicates whether the market is currently set to in-play.

Field	Description	Data Type
id	The unique identifier of the market	float
turnInPlayEnabled	Whether the market is set to turn in-play or not	bool
marketBaseRate	The commission rate applicable to the market	float
eventId	Unique event identifier	int
marketTime	Time	string
suspendTime	The date and time the market is due to be suspended on	string
complete	Whether runners can still be added to the market or not	bool
numberOfActiveRunners	Number of runners currently available to bet on	int

Table 4.5: The final structure of the market definition data

4.2.3 Winners

The file contains the winners and the number of runners in each market. It was extracted from the market definition messages, which originally included the status of each runner as either `ACTIVE`, `REMOVED`, `WINNER`, `PLACED`, `LOSER` or `HIDDEN`. The fields of the winner file are described in table 4.6.

Field	Description	Data Type
id	The unique identifier of the market	float
winner	The identifier of the winning runner	int
eventId	Unique event identifier	int
numberOfRunners	Total number of runners that was available to bet on while the market was open	int

Table 4.6: The final structure of the winners data

4.2.4 Returns

Finally, I calculate and export timestamped raw returns: positive and negative separately as I later compare the distribution of the two. The returns are calculated based on last traded price, the final status of the runner the price was traded for, and the volume traded at that price. Each last traded price signal indicates returns for both the back and lay side, one of them positive one negative. Last traded price represents the decimal odds which are inclusive of stake. Betfair also applies a small commission to profits. The commission rate on Betfair varies depending on a number of factors, however the most common rate is 5% on net winnings. Given the odds x on an outcome $Z \in \{WIN, LOSS\}$, a stake y and a commission rate of c, if a backer backs an event z and the event occurs, the returns on the back bet amount to $x * y$ minus the commission $(x-1) * y * c$, and the returns on the lay bet are equal to $-(x-1)y$. Therefore the net returns total $(1-c)(x-1)y$. In the event of the back bet losing, the returns on the back bet amount to $-y$, while the returns on the lay bet equal y minus the commission $y * c$ and the net returns amount to $-y * c$. The stake y is calculated by subtracting the total volume of trades at a given price on a given runner the last time that price on this runner selection occurred in the records before the current price occurrence from the current one.

I export two files for each for each event: one for positive returns and one for negative. I later combine the positive and negative returns into a total returns file as needed. This is not an issue as the returns contain retain their timestamps. Since the data is divided by event, I included the `marketId` field to separate it by markets later when necessary. It is quicker to filter the data on the spot than produce 7 times more CSV files in order to store data from each market separately. I personally found that storing the data separated into files by event offers the best trade off between the convenience of having small chunks of data ready to load for the purpose of experimentation, initial exploration and implementation testing, and the efficiency of not having to load in too many CSV files, which compared to filtering a DataFrame is a costly operation. All of the returns files follow the same structure, which is shown in table 5.1.

Field	Description	Data Type
ret	Returns	float
t	Time	int
m	Unique market identifier	float

Table 4.7: The structure of the returns data

The returns serve as a base for the input for later analysis. Depending on the particular properties being explored and methods used I used either logarithmic, simple, squared, raw or absolute returns. I would obtain them by simply loading in the files and processing them as I needed them.

4.3 Dataset

The particular set of Betfair historic data used in this study came from February of 2016 horse racing markets. The dataset spans 10 events, 73 markets and 1056766 price increment signals. The tick-by-tick frequency of the data means messages are sent every 50 milliseconds[8].

The mean number of runners in a market is 9.86, with a standard deviation of 4.22, the highest number 21 and the lowest 3. A bet is matched one average every 50 seconds, with the standard deviation of 450.

Chapter 5

Data Analysis

In this chapter I will describe the methodology and process of deriving stylized facts from online sports betting exchange data. I will also present and discuss the obtained results. Further interpretation and critical evaluation of the results follows in the next chapter.

5.1 The distributional properties of returns

This section outlines the exploration of the unconditional distribution of log returns. This includes the presence of heavy tails or lack thereof, the calculation of the tail exponent, and fitting the function that best fits the tails. I also look at the evidence of gain-loss asymmetry by examining the distributions of positive and negative returns separately and comparing them.

Methods

Logarithmic returns can be defined as

$$R_t = ln(P_t) - ln(P_{t-1})$$

where R_t denotes the logarithmic returns from a period of time t, P_t stand for the price from the period t, and P_{t-1} stands for the price from the previous period, $t-1$. Log returns provide an approximation for simple returns and the relationship between the two is shown in 5.1 and 5.2

$$R_t = ln(1 + r_t) \qquad (5.1) \qquad\qquad r_t = exp(R_t) - 1 \qquad (5.2)$$

where R_t denotes log returns and r_t stand for simple returns. Simple returns are defined as

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Logarithmic returns are preferable to simple or raw returns for the purposes of time series data analysis for a number of reasons. Firstly, log returns can be added across time periods, while doing the same with simple returns can lead to misleading results. Simple returns aggregate across assets, log returns aggregate across time. It is more difficult to derive the time series properties of multiplicative processes than additive processes, making log returns more suitable for time series data analysis[14]. Log returns are continuously compounding. This property makes the frequency of compounding irrelevant. Secondly, visualising financial data tends to be easier and graphs tend to be more readable when using logarithmic returns. Log scales allow a large range to be displayed without small values being compressed down into bottom of the graph. Another important property of log returns that makes them more suitable for time series analysis is their symmetry around zero. This means that unlike simple returns, logarithmic returns of equal magnitude but opposite signs will cancel each other out. Finally, the results of standard statistical tests performed on log-transformed data are often no longer relevant to the non-transformed, raw data[33, 46]. Since the stylized facts of the unconditional distribution of financial time series data were most commonly derived from log returns, it only makes sense to use log returns as well, as comparing sports betting exchanges and financial markets is one of the objectives of this project.

When examining the unconditional distribution of returns I use the combined data from all events and markets. There is no reason to separate different markets and taking all of them into account gives stronger results.

5.1.1 The Unconditional Distribution of Returns

I analyse the distribution of returns from 41588 time periods. As shown in table 5.1, the mean of the distribution is close to 0, potentially indicating that the total volume of negative and positive returns are close to being equal. Their distributions might still be different however, therefore the gain-loss asymmetry remains to be explored in section 5.1.2. The standard deviation is 4.0323, making the variation coefficient $c_v = \mu/\sigma$ equal to $c_v = -0.0004$, which is significantly smaller than one empirically found in financial market data by for example Cont[21]. Cont also found negative skewness, which is not the case in Betfair data. However while opposite in sign, the skewness is similar in magnitude. Positive skewness, albeit significantly higher, was also found in prediction market data[75]. This hints at gain-loss asymmetry that is opposite to that commonly found in financial market data.

N. of Observations	Mean	Std. Dev.	Skewness	Kurtosis
41588	-0.0018	4.0323	0.0241	1.0994

Table 5.1: The descriptive statistics of the unconditional distribution of log returns

5.1.2 Tails

The low kurtosis is a somewhat surprising finding. I use the Pearson's definition, where a normal distribution has a kurtosis of 3.0. This means the data is platykurtic and has a broader peak and thicker tails than a normal distribution. The result lies in contrast to what has been known about financial time series for over half a century, namely that the unconditional distribution of financial asset returns has heavy tails. Evidence of heavy tails has been found in prediction markets by Restocchi as well[75]. To further inspect the distribution of the tails I calculate the tail-index α - a parameter determining how heavy the tail is. The higher the α, the thinner the tail. One method of estimating the tail-index of distribution's tails is the Hill estimator[43]. The method is a quasi-maximum likelihood estimator. The estimator's performance depends heavily on the number of tail observations k used. There is a trade-off between the variance and bias of the estimator depending on the value of k. A smaller value reduces the risk of bias and overfitting, especially if the distribution has a smaller sample size or if the tail behavior is unknown. On the other hand, a larger value of k tends to provide a more accurate estimate of the tail index if the distribution has heavy tails. It can also reduce the variability of the estimator and improve its robustness.The choice of k is the topic of an ongoing debate. It is not uncommon to "eye-ball" the appropriate number of tail observations[74], but the most popular approach is choosing the k that minimizes the mean squared error (MSE). Danielsson et al. propose a method based on fitting the tail by minimizing the maximum deviation in the quantile dimension instead[24]. Other estimators include Pickand's tail-index estimator[73] and Ratio estimator[38]. k in table 5.2 takes the form of a proportion of the total number of observations in the distribution.

The values of the Hill estimator for Betfair data are shown in table 5.2. A hill estimator of value -0.3 corresponds roughly a tail-index of 3. The method is however most suitable for heavy-tailed distributions and therefore the result for our data is not necessarily very accurate.

k	Hill estimator
0.01	-0.348
0.02	-0.454
0.03	-0.533
0.04	-0.599
0.05	-0.656
0.1	-0.877

Table 5.2: The values of the Hill estimator

Choosing a Distribution

When fitting a number of distributions to the data the smallest sum squared error was given by a generalized normal distribution with a β parameter equal to 1.19. For $\beta = 1$ the distribution is identical to a Laplace distribution, when $\beta = 2$ the distribution is a Gaussian. The probability density function for a generalized Gaussian is given by

$$f(x, \beta) = \beta/(2\Gamma(1/\beta))exp(-|x|^{\beta})$$

where x is a real number, $\beta > 0$ and Γ is the Gamma function.

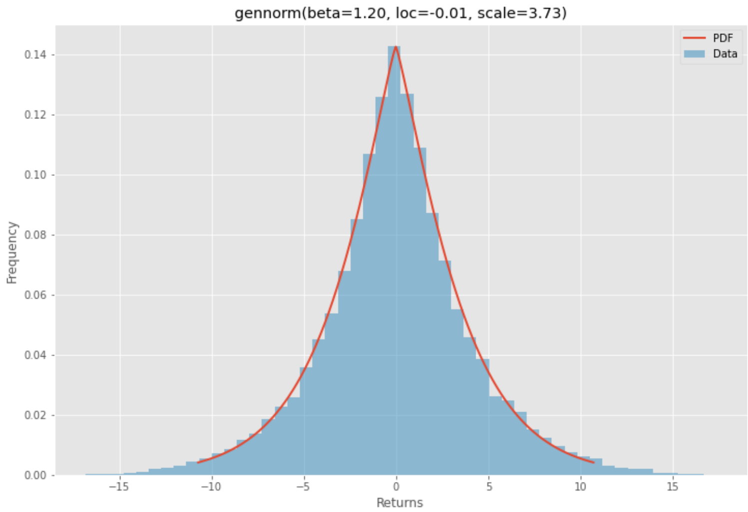

Figure 5.1: A fitting of a generalized normal distribution to the probability density function of the data

5.1.3 Gain-Loss asymmetry

In order to examine the symmetry between negative and positive returns I compare their unconditional distributions. Table 5.3 shows the descriptive statistics of those distributions. The values are very close and the visual inspection of the two distributions as per 5.2 proves them to be nearly identical. To inspect further I perform a two-sample Kolmogorov-Smirnov (KS) test. The null hypothesis H_0 is that the two samples are drawn from the same underlying distribution. The KS test statistic D is the maximum vertical distance between the two empirical distribution functions. The null hypothesis can be rejected if either the p-value is small enough or the value of the statistic D is greater than the critical value D_c defined as

$$D_c = c(\alpha)\sqrt{\frac{n_a + n_b}{n_a n_b}}$$

where $c(\alpha)$ is a constant representing the inverse of the Kolmogorov distribution at a level of significance α (usually assumed to be 0.05) and n_a and n_b are the numbers of observations in the distributions which are being compared. Computing D, D_c and the p-value gives us $D_c = 0.0080$ for $\alpha = 0.05$, $D = 0.0068$ and $p - value = 0.1347$. Since $D < D_c$ and $p > 0.05$, we cannot reject the null hypothesis that the negative and positive returns follow the same distribution. This would indicate the absence of gain-loss asymmetry: the gains and losses are nearly equal in magnitude.

Returns	N. of Observations	Mean	Std. Dev.	Skewness	Kurtosis
Positive	57648	0.0001	3.769	0.0036	1.362
Negative	57648	0.0001	3.628	0.0102	1.325

Table 5.3: The descriptive statistics of the unconditional distributions of positive and negative returns

The distributions that yielded the smallest mean squared errors were generalized normal distributions for both negative and positive returns with values of β close to those found in the distribution of total returns. This is shown in 5.2.

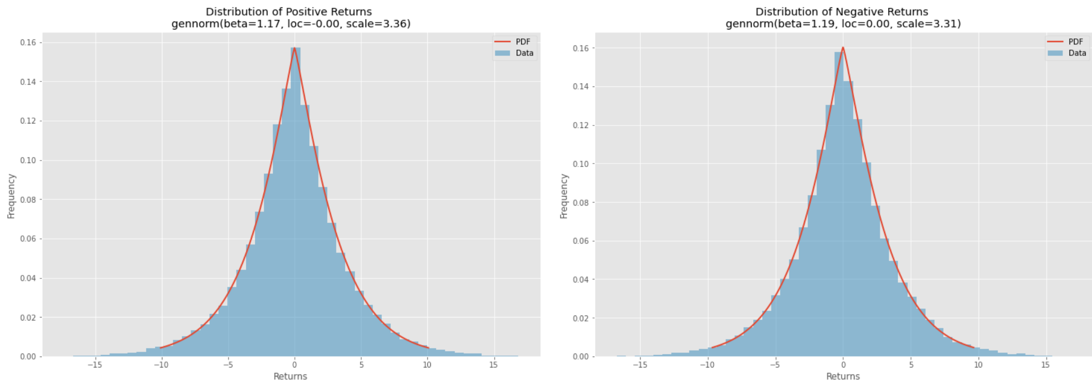

Figure 5.2: A fitting of a generalized normal distribution to the probability density function of positive and negative returns

5.2 The time dependence properties of returns

Time dependence properties of financial time series have been of great interest to researchers[4, 60] as they contribute to the discussion on market efficiency of financial exchanges.

Methods

Unlike for the distributional properties of the data, I employ analysis of not only log returns, but also absolute and squared returns. Those are defined as

$$|R_t| = |P_t - P_{t-1}| \qquad (5.3)$$

$$R_t^2 = (P_t - P_{t-1})^2 \qquad (5.4)$$

5.1 and 5.2 respectively, where $|R_t|$ denotes the absolute and R^2 squared values of returns from a period of time t, P_t stand for price from the period t, and P_{t-1} stands for price from the previous period, $t-1$.

What is more, instead of analyzing the whole dataset at once, I compute the results for each market separately and then summarize them. This choice is motivated by the fact that even if information was somehow carried across multiple markets, the exploration of that phenomenon would lie outside of the scope of this study. Studies on dependence properties of financial time series data usually focus on one market at a time when deriving stylized facts pertaining to time dependence, even though, unlike sports betting markets, the market might run for numerous years, producing a larger amount of data.

5.2.1 Stationarity

For the search for statistical properties of time series data to have any purpose, at least some of those properties have to be stable over time. This condition is satisfied only if the return process in time is stationary. Stationarity ensures one can mix data from different periods. Formally, a stationary signal has an independent joint Cumulative Distribution Function (CDF) under a time-shift. Time series therefore need to be tested for stationarity before applying any analysis aimed at deriving time dependence stylized facts from the data.

Augmented Dickey–Fuller Test

An augmented Dickey-Fuller (ADF) test[26] checks for a unit root in the time series data in the presence of serial correlation. The null hypothesis H_0 is that a series has a unit root and therefore is non-stationary. The alternative hypothesis H_1 states the opposite. The augmented Dickey Fuller statistic is negative number - the lower it is, the stronger the rejection of the null hypothesis. In order to reject the null hypothesis the test statistic has to be smaller than the critical value and the and p-value has to be smaller than 0.05. Table 5.4 presents the results of the ADF test for 3 example markets from the Betfair data. The test was done on absolute returns, as they will later be used to conduct the analysis of other time dependence properties.

Market Id	ADF Statistic	Critical Value (5%)	P-value	Rejected
1.122946937	-5.021294	-2.863	0.000020	True
1.122946927	-8.518810	-2.862	0.000000	True
1.122946942	-5.129096	-2.864	0.000012	True

Table 5.4: The results of the ADF test for 3 example markets

The results for the rest of data were analogous, with very low both ADF statisctics and p-values, meaning we can confidently reject the null hypothesis that the time series has a unit root. The null hypothesis was rejected for all 73 markets. This appears to confirm the trend-stationarity of the data.

Kwiatkowski-Phillips-Schmidt-Shin Test

A Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is a stationarity test. It tests whether the mean and covariance of time series data are statistically time-shift invariant. The null hypothesis H_0 is that the time series is weakly stationary or stationary around a deterministic trend, while the alternative hypothesis H_1 states that the time series has a unit root and is non-stationary. The results on 3 example markets are shown in table 5.5.

Market Id	KPSS Statistic	Critical Value (5%)	P-value	Rejected
1.122946937	1.21043	0.463	0.01	True
1.122946927	2.47760	0.463	0.01	True
1.122946942	0.98900	0.463	0.01	True

Table 5.5: The results of the KPSS test for 3 example markets

It is worth noting that the ADF test and KPSS test are not equivalent and cannot be used interchangeably, which is precisely the reason I chose to employ them both. ADF and KPSS tests have opposite null hypotheses. Generally speaking, a null hypothesis cannot be proven to be true, but it can be confidently rejected, making the alternative hypothesis the stronger one. Thus the two tests complement each other in this trivial way. Furthermore, the difference between the implementation of the tests allows us to make not only stronger, but also more precise observations about the data. If both tests indicate stationarity, we can be more sure of the property and vice versa for non-stationarity. However, if the tests show conflicting results, we can make additional conclusions. Namely if KPSS indicates stationarity and ADF points to non-stationarity the series is trend-stationary. Conversely, when KPSS affirms non-stationarity and ADF indicates stationarity the series is difference stationary.

The results from the KPSS test are opposite to those from ADF. The results from the example 3 markets are representative of the remaining markets: the null hypothesis that the time series is weakly stationary is consistently rejected, again for all 73 markets. As previously described, this result indicates that time series data from online betting exchanges is difference stationary. This means the data would need differenced one or more times to become stationary. Differencing is a process in which the value at the current time step is calculated as the difference between its value and the value at the previous time step. In comparison, having applied the ADF and KPSS tests to log returns they appear to be stationary, as shown in tables 5.6 and 5.7.

Market Id	ADF Statistic	Critical Value (5%)	P-value	Rejected
1.122946937	-13.735484	-2.863	0.00000	True
1.122946927	-20.866090	-2.862	0.00000	True
1.122946942	-10.775032	-2.864	0.00000	True

Table 5.6: The results of the ADF test for 3 example markets for log returns

5.2.2 Autocorrelations

In order to examine temporal dependence and market efficiency it is crucial to analyse the autocorrelations of asset returns. In simple terms, autocorrelation measures the relationship between a variable's current

Market Id	KPSS Statistic	Critical Value (5%)	P-value	Rejected
1.122946937	0.014013	0.463	0.01	True
1.122946927	0.053100	0.463	0.01	True
1.122946942	0.049614	0.463	0.01	True

Table 5.7: The results of the KPSS test for 3 example markets for log returns

and past values.

Linear Autocorrelations

It is a well established stylized fact that high-frequency financial time series data does not exhibit significant linear autocorrelation, except at very short lags where one observes a slowly decaying negative linear autocorrelation, as well as negative autocorrelation at the first lag in the bid or ask price itself. This property can be examined using log returns, which have been shown to be stationary in section 5.2.1.

In Betfair data, log returns appear to follow a similar trend. Upon visual inspection of autocorrelation plots of tick-by-tick returns at the first 30 lags for multiple markets shown in 5.3 one can see that the plots follow a similar shape. One can see the strong negative autocorrelation at the first lag, maybe some at the first few, but after that it is reduced to non-significant levels.

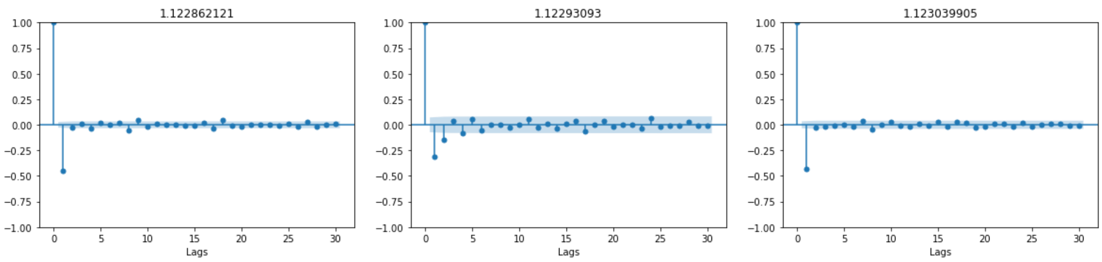

Figure 5.3: Autocorrelation plots for 3 example markets

The lack of significant autocorrelation is a sign of market efficiency. The time it takes to reduce correlations is representative of the time it takes for the market to react to new information, as any significant correlation can be exploited by the traders to conceive profitable betting strategies. If the market is information efficient, the employment of those strategies will quickly nullify the correlation. Whether financial markets are efficient or not is highly contested and has been a topic of study for at least half a century[31]. There is no consesus on the efficient market hypothesis (EMH), however most researchers agree that financial exchanges are at least somewhat effcient[32, 62]. Betting exchange data showing a degree of autocorrelation that is close to that found in financial time series from liquid markets therefore suggests that online sports betting exchanges exhibit a similar level of information efficency.

5.2.3 Volatility Clustering

Volatility clustering is the phenomenon of small changes in price of either sign being followed by small changes and large changes being followed by large changes. This property implies a degree of inefficiency in the market as volatility clustering means the magnitude of the price variations shows a degree of predictability. A common measure of volatility clustering is the autocorrelation function of absolute or squared returns, also known as nonlinear autocorrelation. In financial time series nonlinear functions of returns show a strong positive autocorrelation that decays slowly. This decay has been empirically found to show power-law behaviour with α, the power law exponent, laying between 0.1 and 0.4[58, 21].

The Betfair dataset exhibits significant positive nonlinear autocorrelation. However, when fitted with a power-law it becomes apparent that the autocorrelation function of absolute returns behaves differently than the one in financial market data. As shown in table 5.8 the values of the power-law exponent α are on average higher. This means faster decay and therefore higher market efficiency.

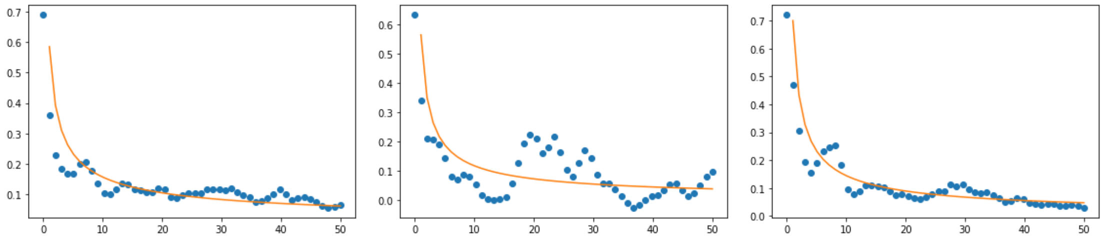

Figure 5.4: Autocorrelation plots for 3 example markets with power-law functions fitted

N. of Observations	Mean	Std. Dev.	Max	Min
73	0.62435	0.3052	1.9755	0.1645

Table 5.8: The values of the power-law exponent α for absolute returns

5.2.4 Long-Range Memory

Long-range memory is a property of a time series where the autocorrelation function decays slowly or not at all, meaning the past changes in volatility impact the volatility far into the future. A random process is said to have long-memory if its autocorrelation function is not integrable. More formally, given the autocovariance function $\gamma(k)$, we define a process as long-memory if in the limit $k \to \infty$ it is true that

$$\gamma(k) \sim k^{-\alpha} L(k)$$

. $L(k)$ is a slowly varying function at infinity. The degree of long-range memory is described by the exponent α which lies in range $\alpha \in (0, 1)$. Smaller alpha corresponds to longer memory and vice versa.

One of the measures of long-range memory is the Hurst exponent, which is related to α in the autocorrelation function by

$$H = 1 - \frac{\alpha}{2}$$

. $0 < H < 0.5$ indicates mean-reversion, meaning the time series values gradually move towards the long-term mean. $0.5 < H < 1$ shows the presence of long-range memory. Finally, $H = 0.5$ indicates a perfect Brownian motion. Brownian motion is a continuous stochastic process characterized by stationary, normally distributed,and independent increments, meaning that the increments of the process over any fixed time interval are independent of one another and have the same distribution. In financial time series data exhibiting Brownian motion the returns would be random and the market perfectly efficient.

Hurst exponent can be estimated in a number of ways. Among the most commonly used methods is the classical R/S test[64] which works by calculating the rescaled range statistic, R/S, which is defined as the ratio of the range of the cumulative deviation of the time series from its mean to the standard deviation of the time series. For long-range memory processes the deviations are larger. It has however been criticised for being too weak, meaning it can falsely claim the series exhibits long-memory when it does not, as shown by Lo[59]. Furthermore, it is based on assumptions that do not necessarily hold true for sports betting exchange time series data, namely that the time series is stationary and Gaussian. In the same paper, Lo introduced a modified version of the test. It has however been proven to often be too strong as it assumes that the sample size of the time series is large enough to obtain accurate estimates of the Hurst exponent[89].

The values of the Hurst exponent estimated via classical R/S analysis on Betfair data are shown in table 5.9. I use log returns, as their distribution is the closest to a Gaussian and they are stationary. As clearly visible from the table, the value of the exponent depends heavily on the number of lags. The appropriate number is a topic of debate, but some of the approaches include choosing one-third to one-half of the length of the time series. The values of H rise with the number of lags. The number of observations in this case is 41588, therefore for any reasonable number of lags (smaller than half of the sample size) the values of the exponent suggest mean reversion. This does somewhat align with the previous results concerning the rate of decay for nonlinear autocorrelations, which is significantly higher in betting exchange data than it is typically in financial time series. Considering that the classical R/S test is unlikely to return false negatives when it comes to the presence of long-range memory, this result is relatively significant despite the test's unreliability.

N. of Lags	H
20	0.0003
100	0.0013
300	0.0033
500	0.0046

Table 5.9: The values of Hurst exponent estimated via classical R/S analysis depending on the number of lags used

The markets on online betting exchanges are magnitudes shorter-lasting than financial markets, therefore it needs to be stated that the term long-range will not have the same meaning as in the case of financial exchanges. While there has been evidence of long-range memory over months or even years in financial time series, Betfair horse-racing markets typically only last a few hours up to a day.

Chapter 6

Discussion

This chapter serves as a further examination of the obtained results, as well as some of their limitations. I also aim to propose causes and explanations wherever possible and critically assess the validity of the results.

6.1 Lack of heavy tails

The first of the observations made in this study I would like to investigate further is the lack of heavy tails in the unconditional distribution of log returns. The result is unexpected as it stands in contrast to the one of the most well established stylized facts of financial time series. One could theorize that perhaps sporting events, especially horse racing, are somewhat predictable. The extreme events that cause high volatility in financial markets and thus contribute to their heavy-tailedness simply do not occur. What kind of event during a horse race could possibly be analogous to the kind of world-changing events that influence stock markets? Political prediction markets are often influenced by the kind of events financial markets are, which would explain why Restocchi found heavy tails in prediction market data as well while I have not. Another possible explanation could be higher market efficiency, which I will expand on in the following section. Higher efficiency means fewer and more quickly corrected anomalies, which would make the distribution of returns have lighter tails.

6.2 Informational efficiency

Nonlinear autocorrelations in sports betting exchange data showed faster decay than they do in financial time series. This, together with evidence of mean reversion and lack of significant linear autocorrelations suggest market efficiency. This is in line with the existing research on online sports betting exchange data summarised in section 3.2.2. Those results however are not fully conclusive, as fitting the power law to the autocorrelation function of absolute returns could use more data for improved accuracy and so would the Hurst exponent estimation. Furthermore, clasical R/S analysis is known to be weak and more informative results would perhaps be obtained if Detrended Fluctuation Analysis was implemented instead. However, if we do accept the conclusion that sports betting exchanges are more informationally efficient than financial markets it could potentially be interpreted in a number of ways. One possible explanation could be that as previously mentioned in section 2.1.3, sports betting exchanges, as opposed to traditional bookmakers, attract professional or sophisticated bettors. As shown by Abinzano, Muga and Santamaria[1], small bettors tend to be the ones generating mispricing. The higher number of sophisticated traders might correspond to higher market efficiency.

6.3 Gain-Loss symmetry

A two-sample Kolmogorov-Smirnov test on the positive and negative log returns has indicated they come from the same distributions. Fitting distributions to both seemed to confirm that finding. This implies the lack of the gain-loss asymmetry property in the distribution of returns from betting exchanges. The most trivial explanation could be the length of the period the data covers. It might have simply not been long enough to capture the asymmetry as it may vary across time. An opposing theory would be that the gain-loss asymmetry is not present in betting exchange time series in general and that it is caused by betting exchanges attracting less risk averse participants than financial markets and thus are less put off by losses. While there is evidence of betting exchange users being more risk averse than those who bet with traditional bookmakers[69], I am not aware of any studies comparing them with stock traders.

6.4 Calendar effects

One of the caveats concerning the extent and applicability of the results is one that is a concern for all research involving data analysis. That is whether the dataset was big enough an whether it can be considered representative of the subject being studied. As previously described, the data for this study spans a month of trading activity of Betfair. Meanwhile numerous studies on financial markets examine time series that last many years[17, 48]. Besides larger amounts of data increasing the accuracy of fitting any functions or distributions, as well as the strength of the statistical tests applied, there is no guarantee that a month is representative of the whole year. There is no reason why betting exchange data could not exhibit a month-of-the-year effect known from financial time series[53]. That would mean that analysing a different moth could bring different results and the only way to know would be using data from a significantly longer time period.

Chapter 7

Conclusions

The project was successful in identifying a range of stylized facts in online sports betting exchanges by conducting an analysis of high-frequency time series data from Betfair horse racing markets. Having explored the unconditional distribution of returns, evidence of light tails was found in the form of a tail index closer to that of a Gaussian than a power law. Statistical testing and analysis have also shown the lack of gain-loss asymmetry in this distribution. The absolute returns were concluded to be difference stationary and log returns stationary. Linear autocorrelations proved to be negligible except for the first few lags and nonlinear autocorrelations showed faster decay than in financial time series, indicating a lesser degree of volatility clustering. The absence of long-range memory in betting exchange time series was affirmed by the estimation of the Hurst exponent, which pointed to evidence of mean reversion. Overall the results suggest higher informational efficiency than that typically exhibited by financial markets according to the stylized facts about them.

Deriving stylized facts is by definition an open ended question. One cannot confidently say they have explored something so thoroughly no more new properties can be identified in it, especially when talking about a thing as complex and dynamic as betting markets. Therefore for this project there was no set finite list of statistical properties whose presence or absence had to be explored. The general plan however was to investigate if the most established and famous stylized facts of financial time series also hold true in online sports betting exchanges. It offers a strong starting point for research on stylized facts in online sports betting exchanges as they do bear resemblance to financial markets and that similarity is often the motivation for studying betting exchanges in the first place. Another factor I took into account was whether the properties I was about investigate had been studied before. I was not interested in exploring the favourite-longshot bias for example as it had already been extensively studied before. Additionally, one could dedicate their entire thesis just to this effect. There were a few points in that plan that I did not manage to cover due to time constraints as well. Those include leverage effect and aggregational gaussianity. For the topics that I did cover, I would like to employ some more advanced analysis techniques. I would use Detrended Fluctuation Analysis instead of classical R/S to estimate the Hurst exponent. The only reason I did not was due to issues with implementation and again, time constrains.

7.1 Further Work

This leads me to the next point, which is the further work this project would benefit from. As outlined in the previous paragraph, there is a number of stylized facts in financial time series whose presence in online sports betting exchange data could still be explored, such as conditional heavy tails, volume/volatility correlation or asymmetry in time scales. Furthermore, this analysis could be done on more data and include other sports, not exclusively horse racing. Betting markets for different disciplines could as well display completely different characteristics. Another compelling avenue for future research would be exploring how those statistical tendencies differ between markets that are currently in play and those which are not. There is a high chance of seasonal effects existing in betting exchange data, except relating to events such as the market going in-play instead of actual calendar events. Finally, the stylized facts could be reproduced by a synthetic generator of betting exchange data, like the Bristol Betting Exchange[20], which was one of the original motivations for the project.

Bibliography

[1] Muga L. Santamaria R. Abinzano, I. The role of small bettors in price formation in betting exchanges. *Applied Economics*, 2020.

[2] Muga L. Santamaria R. Abínzano, I. Game, set and match: the favourite-long shot bias in tennis betting exchanges. *Applied Economics Letters*, 1-4, 2015.

[3] Bollerslev T. Diebold F.X. Andersen, T.G. and P. Labys. Modeling and forecasting realized volatility. *Econometrica*, 71, 2003.

[4] Bollerslev T. Diebold F.X. Ebens H. Andersen, T.G. The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1), 2001.

[5] J. P. Aruajo-Santos. A trading agent framework using plain strategies and machine learning. 2014.

[6] G. Axen and D. Cortis. Hedging on betting markets. risks. 2020.

[7] PA Bebbington. Studies in informational price formation, prediction markets, and trading. 2017.

[8] Betfair. *[online] Available at: https://historicdata.betfair.com/home [Accessed 2 Apr 2023].*

[9] Kiesel R. Bingham, N. *Risk-neutral valuation: Pricing and hedging of financial derivatives*. Springer-Verlag, 2004.

[10] F. Black. Studies of stock price volatility changes. *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Statistics Section*, 1976.

[11] T. Bollersev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 1986.

[12] DOMOWITZ I. BOLLERSLEV, T. Trading patterns and prices in the interbank foreign exchange market. *The Journal of Finance*, 1993.

[13] R. Bunker and T. Susnjak. The application of machine learning techniques for predicting results in team sport: A review. 2019.

[14] A. W. Lo Campbell, J. Y. and A. C. Mackinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

[15] Masanell Casadesus. Platform competition: Betfair and the uk market for sports betting. *Journal of Economics and Management Strategy*, 2019.

[16] Campbell N. Casadesus-Masanell, R. Platform competition: Betfair and the uk market for sports betting. *Journal of Economics and Management Strategy*, 28, 2019.

[17] Darné O. Charles, A. Large shocks in the volatility of the dow jones industrial average index. *Journal of Banking Finance*, 43, 2012.

[18] M. et al. Chegere. It's a sure win! experimental evidence on overconfidence in betting behavior. *CSEF Working Papers*, 655, 2022.

[19] Joseph S. Chen, Harrison G. Hong, and Jeremy C. Stein. Forecasting crashes: Trading volume, past returns and conditional skewness in stock prices. 2000.

[20] D. Cliff. Bbe: Simulating the microstructural dynamics of an in-play betting exchange via agent-based modelling. 2021.

[21] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantative Finance*, 1, 2001.

[22] Reade J. J. Croxson, K. Exchange vs. dealers: A high-frequency analysis of in-play betting prices. 2010.

[23] Ramazan; Müller Ulrich; Olsen Richard B. Dacorogna, Michel M.; Gençay. *An Introduction to High Frequency Finance*. Academic Press, 2001.

[24] Ergun L. M. de Haan L. de Vries-C. G. Danielsson, J. Tail index estimation: Quantile-driven threshold selection. 2019.

[25] Rindi B. de Jong, F. *The Microstructure of Financial Markets*. Cambridge University Press, 2009.

[26] D. Dickey and W. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 1979.

[27] Granger C. W. J. Engle R. F. Ding, Z. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1, 1993.

[28] Keller U. Eberlein, E. Hyperbolic distributions in finance. bernoulli. 1995.

[29] R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4), 1982.

[30] Braverman et al. How do gamblers start gambling: identifying behavioural markers for high-risk internet gambling. *European Journal of Public Health*, 2013.

[31] E. F. Fama. The behavior of stock-market prices. *The Journal of Business*, 38(1), 1965.

[32] E. F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2), 1970.

[33] Lu N Chen T He H-Lu Y Tu XM Feng C, Wang H. Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry*, 2, 2014.

[34] M. Flood. Market structure and inefficiency in the foreign exchange market. *Journal of International Money and Finance*, 13, 1994.

[35] Verbeek E. Nuesch S. Franck, E. Prediction accuracy of different market structures — bookmakers versus a betting exchange. *International Journal of Forecasting*, 26, 2010.

[36] GitHub. *[online] Available at: https://github.com/betcode-org/betfair [Accessed 2 Apr 2023]*.

[37] Jagannathan R. Runkle. D. E. Glosten, L. R. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48, 1993.

[38] Smith R.L. Goldie C.M. *Slow variation with remainder: theory and applications.*, volume 38. Quart. J. Math. Oxford, 1987.

[39] GrandViewResearch. [online] available at: https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report [accessed 16 mar 2023].

[40] LaPlante D. A. Shaffer H. J. Gray, H. M. Behavioral characteristics of internet gamblers who trigger corporate responsible gambling interventions. *Psychology of Addictive Behaviors*, 26, 2012.

[41] Auer M. Griffiths, M. The irrelevancy of game-type in the acquisition, development, and maintenance of problem gambling. *Sec. Psychopathology*, 3, 2013.

[42] Richmond P. Hutzler S. Hardiman, S. Long-range correlations in an online betting exchange for a football tournament. *New Journal of Physics*, 12, 2010.

[43] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3, 1975.

[44] Smith M. Rockloff M. et al. Hing, N. How structural changes in online gambling are shaping the contemporary experiences and behaviours of online gamblers: an interview study. *PBMC Public Health*, 22, 2022.

[45] H. S. Houthakker. Systematic and random elements in short-term price movements. *The American Economic Review*, 51, no. 2, 1961.

[46] R. Hudson and A. Gregoriou. Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. 2010.

[47] Johansen A. Simonsen I. Jensen, M. H. Inverse statistics in economics: the gain–loss asymmetry. *Physica A: Statistical Mechanics and its Applications*, 324, 2003.

[48] Lee J. Johnson, T. C. On the systematic volatility of unpriced earnings. *Journal of Financial Economics*, 114, 2014.

[49] D. Kahneman and A. Tversky. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 1992.

[50] K. J. Kain and T. D. Logan. *Are Sports Betting Markets Prediction Markets?: Evidence From a New Test*. 15(1) edition, 2014.

[51] A. Kan. Efficiency, bias, and decisions: Observations from a sports betting exchange. 2015.

[52] M. G. Kendall. The analysis of economic time-series—part i: Prices. *Journal of the Royal Statistical Society*, 116(1), 1953.

[53] A. Klesov. Calendar effects on stock market: Case of selected cis and cee countries. 2008.

[54] Velzen B. Koning, R. Betting exchanges: the future of sports betting? *International Journal of Sport Finance*, 4, 2009.

[55] Zijm R. Koning, R.H. Betting market efficiency and prediction in binary choice model. *Ann Oper Res*, 2022.

[56] Trevon D. Logan Kyle J. Kain. Are sports betting markets prediction markets?: Evidence from a new test. 2014.

[57] Lee Jun-De Lee Chi-Chuan Lee, Chien-Chiang. Stock prices and the efficient market hypothesis: Evidence from a panel stationary test with structural breaks. *Japan and the World Economy*, 22, 2010.

[58] Gopikrishnan P. Cizeau-P. Meyer M.-Peng C.-K. Liu, Y. and H. E. Stanley. Statistical properties of the volatility of price fluctuations. *Physical Review*, 60(2), 1999.

[59] A.W. Lo. Long-term memory in stock market prices. *Econometrica*, 59, 1991.

[60] MacKinlay A.C. Lo, A.W. Stock market prices do not follow random walks: Evidence from a simple specification test. 1988.

[61] H. Lopez-Gonzalez and M.D. Griffiths. Understanding the convergence of online sports betting markets. *International Review for the Sociology of Sport*, 2016.

[62] B. G. Malkiel. The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, 17(1), 2003.

[63] B. Mandelbrot. The variation of certain speculative prices. *Journal of Business*, XXXVI, 1963.

[64] B. Mandelbrot. Statistical methodology for nonperiodic cycles: From the co- variance to r/s analysis. *Annals of Economic and Social Measurement*, 1, 1972.

[65] R.; Franck E Meier, P. F.; Flepp. Are sports betting markets semistrong efficient? evidence from the covid-19 pandemic. 2021.

[66] F. Mittelbach, M. Goossens, J. Braams, D. Carlisle, and C. Rowley. Time-varying periodicity in intraday volatility. 2018.

[67] H. Nyberg. Qr-garch-m model for risk-return tradeoff in u.s. stock returns and business cycles. 2011.

[68] M. O'Hara. *Market Microstructure Theory*. Wiley, 1998.

[69] M. Ottaviani and P. N. Sørensen. Price reaction to information with heterogeneous beliefs and wealth effects: Underreaction, momentum, and reversal. *American Economic Review*, 105(1), 2015.

[70] L. D. Pankoff. Market efficiency and football betting. *Journal of Business*, 1968.

[71] Cain M. Law D. Peel, D. The favourite-longshot bias and market efficiency in uk football betting. *International Journal of Political economy*, 47, 2000.

[72] A. Peiro. Skewness in financial returns. *Journal of Banking and Finance*, 23, 1999.

[73] J. III Pickands. Statistical inference using extreme-order statistics. *The Annals of Statistics*, 19, 1975.

[74] Stărică C. Resnick, S. Smoothing the hill estimator. *Advances in Applied Probability*, 29(1), 1997.

[75] V. Restocchi. It takes all sorts: the complexity of prediction markets. *University of Southampton*, 2018.

[76] VICTOR K. NG ROBERT F. ENGLE. Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48, 1993.

[77] R. Sauer. The economics of wagering markets. *Journal of Economic Literature*, 1998.

[78] M. Shakeel and B. Srivastava. Stylized facts of high-frequency financial time series data. *Global Business Review*, 22(2), 2021.

[79] Similarweb. [online] available at:https://www.similarweb.com/website/betfair.com/ [accessed 15 mar 2023].

[80] Paton L. Vaughan Williams Smith, D. Market efficiency in person-to-person betting. *Wiley*, 2006.

[81] R. Solow. *Growth Theory: an Exposition*. Oxford Univ. Press, 1969.

[82] N. Suhonen. Normative and descriptive theories of decision making under risk: A short review. 2007.

[83] Shaw R. Swindler, S. Racetrack wagering and the "uninformed" bettor: A study of market efficiency. *The Quarterly Review of Economics and Finance*, 35, 1995.

[84] S. J. Taylor. *Modelling Financial Time Series*. Wiley, 1986.

[85] S. J. Taylor. Asset price dynamics, volatility, and prediction. *Princeton University Press*, 2005.

[86] R. S. Tsay. Analysis of financial time series. 2004.

[87] S. B. Wachtel. Certain observations on seasonal movements in stock prices. *The Journal of Business*, 16, 1942.

[88] Wikipedia. Betfair [online] available at: https://en.wikipedia.org/wiki/betfair [accessed 13 mar. 2023].

[89] Taqqu M.S. Willinger, W. and V. Teverovsky. Stock market prices and long-range dependence. *Finance and Stochastics*, 3, 1999.

[90] R. T. Wood and R. J. Williams. Internet gambling: Prevalence, patterns, problems, and policy options. 2009.

Document title:	diss_upd.pdf
Filename:	diss_upd.pdf.pdf
Document id:	6bbea64c-4207-450b-bad0-d4102613585a
Document status:	Completed

Document Audit Trail

CREATED AT: May 04, 2023 | 09:37:13 UTC+00:00

Document Created

UPDATED AT: May 04, 2023 | 09:39:59 UTC+00:00

Document updated

SIGNED AT: May 04, 2023 | 09:40:01 UTC+00:00

Signed by Me (Now) | k.kadzielawa00@gmail.com

From IP: 86.7.192.160

Document Audit Trail

COMPLETED AT: May 04, 2023 | 09:40:01 UTC+00:00

Document Completed