Predictive Modelling for Football Betting Strategies

by

Charaka Abeywickrama

Candidate Number: ZGDP0

MSc Data Science & Machine Learning

Under the supervision of

Prof. Philip Treleaven

September 2023

Department of Computer Science

University College London

Declaration

I Charaka Abeywickrama, declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included and referenced. The report may be freely copied and distributed provided the source is explicitly acknowledged.

C Abeywickrama	25/09/22
Signature	*Date*

Abstract

This thesis investigates the application of machine learning models and betting strategies to predict outcomes in football matches and contribute to a pioneering algorithmic betting platform. This research is propelled by the hypothesis that football match outcomes and in-play events reveal distinct patterns and trends that can be effectively leveraged for predictive modelling. This study designs and implements machine learning models that predict match results based on pre-match information, as well as in-play momentum using betting exchange data. The thesis comprises of two investigations, in the field of sports analytics and machine learning. The structure of the research is presented as follows:

Experiment 1: *Value Betting Strategy Framework in Football*. This study presents an exploration into the use of machine learning models for predicting football match outcomes. The research unfolds a detailed framework for constructing, evaluating, and refining these models, alongside correlating value betting strategies. It also includes a detailed process for the synthesis of novel feature sets from pre-match data, enhancing the prediction capabilities of the models. Central to the discussion is the formal definition of a classification problem for machine learning algorithms designed to model football match outcomes. The study iteratively builds from simple fixed betting strategies to more advanced strategies, leveraging the predictions from the machine learning models. The models' effectiveness is enhanced by the synthetic features, providing a more nuanced understanding of match outcomes.

Experiment 2: *Momentum Betting Using In-Play Data*. The study introduces the construction of an in-play momentum betting strategy utilising stacked LSTM networks, inspired by techniques used in algorithmic trading. The discussion presents betting exchange data and derived features under various in-play scenarios, whereby changes in match momentum give rise to different betting opportunities. The insights are used to develop deep learning models that make inferences on real-time momentum changes and devise effective betting strategies accordingly.

The thesis presents the following original contributions to the field of sports analytics and machine learning:

1. **In-Play Momentum Betting using Stacked LSTM Networks:** The investigation of in-play momentum betting using stacked Long Short-Term Memory (LSTM) networks is an application of deep learning techniques in the realm of football betting. This contribution extends the growing body of research on the application of deep learning in sports analytics.

2. **Robust Backtesting Framework and Data Listener for In-Play Football Data:** The creation of a robust backtesting framework and a data listener for in-play football data provides a practical tool for collecting and analysing in-play data. This facilitates further research and the development of in-play betting strategies, making it a valuable tool for researchers and practitioners.
3. **Comparative Analysis of Value Betting and Momentum Betting Strategies:** A comparative analysis of value betting and momentum betting strategies provides insights into their relative strengths and weaknesses. This analysis sheds light on the strategies' suitability for different types of matches and market conditions, contributing to a more nuanced understanding of their practical application.

Through these contributions, this thesis advances the frontier of research in sports analytics and machine learning, and provides practical tools and strategies for football betting.

Impact Statement

By intertwining advanced machine/deep learning techniques with the world's most popular sport, this thesis not only pushes the boundaries of sports analytics research but also holds immense promise for the multi-billion-dollar football betting industry. Here's an outline of the impacts this research can potentially have on academia, industry, and the wider community:

1. **Setting a New Standard in Predictive Modelling for Football**: By establishing a comprehensive framework, this research sets an example for constructing, evaluating, and refining machine learning models specific to football match prediction. This not only benefits future academic endeavors in the space but also offers clubs, coaches, and players a fresh perspective to anticipate match outcomes, refine strategies, and thereby enhance gameplay.
2. **Practical Tools for Betting and Collecting Data**: The introduction of a robust back testing framework and a data listener for in-play football data represents a tangible asset. These tools not only bridge the gap between research and real-world application but also cater to the needs of modern-day betting platforms seeking to augment their in-play betting strategies.
3. **Holistic View on Betting Strategies**: The comparative analysis between value betting and momentum betting sheds light on the nuances of each, guiding stakeholders in selecting the strategy best suited to specific matches and market dynamics.

The conclusions from these findings could assist everyone from academic researchers to professional gamblers, and even to casual fans looking for an analytical edge in understanding the beautiful game.

Acknowledgements

I would like to express my deepest gratitude to my academic supervisor, Prof Philip Treleaven, for his support and guidance throughout this project. His insightful feedback and encouragement have been invaluable in shaping this research. I am equally thankful to my industry supervisor, John Goodacre from Quant Sports Trading Ltd, for his practical advice and industry insights that have significantly enriched this study. His passion for the field has been truly inspiring.

A special note of thanks goes to UCL for providing an excellent academic and research environment. The resources and facilities available here have been instrumental in the completion of this project.

Lastly, but certainly not least, I would like to thank my family and friends for their unwavering belief in me and for their patience and understanding during this journey.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

The purpose of this introductory chapter is to provide a comprehensive overview of the thesis. The chapter commences with an introduction to the domain of sports betting and predictive modelling, underlining the criticality of comprehending the factors that influence football match outcomes, which in turn contribute to successful betting strategies. The chapter proceeds to outline the specific objectives, experiments, and contributions of this work. It concludes with an overview of the thesis structure, thereby setting the stage for the ensuing in-depth exploration of each individual chapter.

This thesis delves into the world of predicting the outcomes of football matches and devising betting strategies based on these predictions. It examines factors related to football matches, such, as team performance metrics, home advantage and bookmaker odds to assess their power. Additionally it explores betting strategies to determine their effectiveness. The ability to accurately predict match outcomes and develop betting strategies is essential for maximizing returns and minimizing risks in the field of sports betting making this study highly relevant for both researchers and industry practitioners.

The first part of this study involves creating a set of features using pre match data. It also focuses on developing machine learning models that can effectively predict match outcomes. In the part these predictions are used as a foundation for designing betting strategies ranging from simple fixed bets to more complex value betting approaches. A considerable portion of the research is dedicated to in play betting utilizing data from betting exchanges to predict, in play events and formulate momentum based betting strategies. Additionally efforts are made towards developing frameworks that enable researchers to quickly build, test and enhance models and strategies.

Throughout the research process strict academic oversight was maintained to ensure the validity and reliability of the findings. The findings of this thesis offer insights, into the prediction of football outcomes. Betting, while also establishing a foundation for future exploration in sports analytics and machine learning. These research outcomes have implications not for academia but also for individuals involved in sports betting, data analysis and football enthusiasts. This confirms the significance and interest, in this field of study.

1.1. MOTIVATION

The intricate and unpredictable nature of football matches has always intrigued enthusiasts and analysts alike. While there exists a substantial body of research on forecasting football match outcomes based on historical and pre-match data, there remains a significant gap in understanding the dynamics of in-play events. This gap becomes even more pronounced when one considers the vast betting market surrounding football matches. Predicting match outcomes and leveraging them for betting strategies isn't merely an academic endeavor; it has vast economic implications. The betting market is immense, and the ability to accurately predict outcomes can lead to significant financial gains.

Our motivation stems from a desire to bridge this knowledge gap. The unpredictable momentum shifts during a game, influenced by events such as goals, fouls, and player substitutions, can provide a goldmine of betting opportunities. Drawing parallels with the financial markets, where algorithmic trading strategies capitalize on market momentum, we see an opportunity to harness these momentum shifts in football matches using advanced machine learning models.

Furthermore, the world of football betting is vast and multifaceted, with both pre-game and in-play betting holding significant positions in the betting landscape. Pregame betting, which relies on predictions made before the match's commencement, is often rooted in historical data, team performance metrics, and other known factors leading up to the game. In contrast, in-play betting is dynamic, capitalizing on the ever-changing landscape of a live match, where momentum shifts, unexpected events, and split-second decisions can alter the course of the game. While there exists substantial research and strategies centered around pre-game betting, the volatile nature of in-play betting remains less charted.

Additionally our motivation also stems from a keen interest in understanding the nuances and effectiveness of both these betting avenues. By examining both pre-game and in-play betting, we aim to discern the inherent advantages and challenges each presents, and how machine/deep learning can be harnessed to optimize strategies in both domains.

In essence, our research motivation is a blend of academic curiosity, a desire to innovate in the domain of sports analytics, and the economic potential of devising effective betting strategies. Through this research, we aspire to navigate the complexities of football match dynamics, uncover patterns, and devise strategies that are both insightful and profitable.

1.2. RESEARCH OBJECTIVES

The main goal of this research is to investigate and create methods, for predicting the outcomes of football matches using machine learning models and strategies for in play betting. In the phase of the study our focus is on generating features based on pre match data to train various machine learning models. The aim is to achieve accuracy in predictions and gain insights into match outcomes.

Additionally we apply these models to develop betting strategies that offer value starting with simple fixed betting approaches and gradually introducing more sophisticated strategies based on the models predictions.

An important aspect of this research involves designing and implementing a momentum based in play betting strategy, which's relatively unexplored in sports analytics. Taking inspiration from algorithms used in trading this strategy utilizes stacked LSTM (Long Short Term Memory) networks to analyze data from betting exchanges and track the momentum of a game in time. The objective is to identify moments during a match where shifts in momentum present betting opportunities. We also aim to establish a framework for testing improving and comparing strategies.

Lastly an essential objective of this research is to draw comparisons between value based betting strategies and momentum based, in play betting. Each of these approaches has its strengths and weaknesses. Gaining insights, into their performance in various market conditions can be valuable. The purpose of this analysis is to explore the potential of combining these strategies to manage risks and maximize returns.

In general this research aims to broaden our knowledge about the use of machine learning, in sports analytics and betting tactics. By doing it seeks to offer tools and frameworks for football betting while contributing to the advancement of this field.

1.3. RESEARCH EXPERIMENT

This thesis embarks on a detailed examination of predicting football match outcomes and devising betting strategies based on these predictions. The exploration begins with a broad analysis of football match outcome prediction in the first experiment, delves deeper into the realm of betting strategies in the subsequent experiment, and concludes with an in-depth analysis of in-play betting using betting exchange data. The research is divided into two comprehensive studies:

1. **Value Betting Strategy Framework in Football**: The main goal of the study is to gain an understanding of how football match results unfold and to develop strategies, for making valuable bets based on those outcomes. To achieve this the research examines theories from existing literature, such as the connection between team performance metrics and match results well as the influence of playing at home. In contrast to much of the research that usually focuses on aspects of match outcomes this study takes a more comprehensive approach by analyzing a wide range of factors. The analysis also. Validates theories related to finding value in betting on football matches, including examining the effectiveness of the Kelly Criterion. Conducting an analysis within this sample is essential for comprehending the factors that influence match results and for creating betting strategies. The study employs machine learning techniques like feature selection, with validation and classification models.

2. **Momentum Betting Using In-Play Data**: This research project explores how machine learning models can be used to predict events that happen during a game and create betting strategies based on those predictions. We. Analyze data from, in play betting exchanges. Our study is one of the first to apply stacked LSTM networks, which are influenced by techniques used in algorithmic trading to football betting. Most existing studies focus on predicting outcomes before a match begins and do not consider the nature of events that occur during the game. Our research aims to bridge this gap by providing insights into how in play events relate to opportunities for betting. We investigate whether changes in momentum throughout a match can be effectively utilized for in play betting decisions as exploring the impact of different in play events, on these opportunities.

The objective of these experiments is to contribute to the field of football analytics, machine learning and the development of betting strategies. The results of these experiments will offer insights and tools, for predicting football outcomes and making bets while also setting the stage for further research, in this area.

All the code written for these experiments is stored in three GitHub repositories. For Experiment 1, visit here. Experiment 2's code can be accessed here. Finally, the listener code is available here.

1.4. CONTRIBUTIONS TO SCIENCE & INDUSTRY

In the ever-evolving landscapes of science and industry, research endeavors often bridge the gap between theoretical knowledge and practical implementation. The contributions stemming from this study aim to do this, by providing both a deeper scientific understanding and tangible tools for the industry.This thesis has yielded several contributions, each of which is outlined below:

1. **An Extensive Framework for Predictive Modelling in Football:** At the heart of this thesis lies the development of a meticulous framework tailored for predictive modelling in football. This framework, grounded in rigorous machine learning methodologies, serves as a blueprint for constructing, evaluating, and iterating on prediction models. Its systematic nature ensures scalability and adaptability, positioning it as an invaluable asset for future academic explorations and real-world applications in the realm of football predictions.
2. **Exploration and Comparison of Various Betting Strategies:** Venturing beyond mere predictions, this research delved deep into the world of betting strategies. Through extensive experimentation and analysis, a spectrum of strategies, from simple to sophisticated, was explored. The comparative analysis brought forth nuanced insights, revealing the interplay between the strategies and their corresponding predictive models. This exploration provides a roadmap for selecting and implementing betting strategies that align with specific predictive models, optimizing potential returns.
3. **In-Play Momentum Betting using Stacked LSTM Networks:** Marking a significant departure from traditional techniques, this research introduced

the innovative application of stacked Long Short-Term Memory (LSTM) networks for in-play momentum betting. By harnessing the power of deep learning, the research bridged the gap between advanced machine learning and sports betting, paving the way for future innovations in this interdisciplinary domain.

4. **Robust Backtesting Framework and Data Listener for In-Play Football Data:** Recognizing the importance of real-time data in the dynamic world of in-play betting, the thesis introduced a robust backtesting framework complemented by a state-of-the-art data listener. This dual toolset not only facilitates the seamless collection and analysis of in-play football data but also ensures that betting strategies are tested and refined in a realistic environment. This contribution stands as a testament to the practical orientation of the research, underscoring its value for both researchers and industry practitioners.
5. **Comparative Analysis of Value Betting and Momentum Betting Strategies:** In a bid to provide a holistic view of the betting landscape, the research undertook a comprehensive comparative analysis between value betting and momentum betting strategies. This analysis demystified the strengths, limitations, and applications of each strategy, shedding light on their suitability across diverse match scenarios and market dynamics. The insights derived from this analysis offer a nuanced understanding, guiding punters in making informed betting decisions.

The contributions outlined above not only enrich the academic discourse in sports analytics and machine learning but also hold significant practical implications, especially for stakeholders in the sports betting industry.

1.5. THESIS STRUCTURE

The structure of this thesis is organised as follows:

- **Chapter 2 – Background and Literature Review.** In this chapter we will explore the literature and key ideas related to sports analytics, machine learning and sports betting. Our goal is to give you an overview of the issues and context discussed in this thesis. Firstly we will discuss the landscape of football match prediction and betting. Then we will delve into prediction models and strategies used in this field. Additionally we will review the methodologies of machine learning techniques and statistical analysis methods that're relevant, to our research. We'll specifically focus on how these approaches address data analysis challenges, in our study.
- **Chapter 3 – Value Betting Strategy Framework in Football.** In this chapter we will delve into the literature and important ideas surrounding sports analytics, machine learning and sports betting. Our goal is to give you an introduction, to the topics addressed in this thesis. We'll begin by providing an overview of the context of football match prediction and betting. Then we'll dive into a discussion on prediction models and strategies for betting. Additionally we'll explore machine learning techniques and

statistical analysis methods that have been used to tackle data analysis problems, in research.

- **Chapter 4 – Momentum Betting and Drift Betting.** In this chapter we delve into the experiment of the thesis that centers around, in play betting. We dive into how we utilize data from betting exchanges to forecast in play events and craft momentum based betting models. Additionally we explore the implementation of stacked LSTM networks drawing inspiration from techniques commonly utilized in trading. The chapter encompasses an explanation of our methodology, an examination of the results obtained and an evaluation of the developed, in play betting strategies.
- **Chapter 5 – Conclusion.** In this concluding chapter we will provide an overview of the research undertaken highlighting the contributions this thesis has made to the fields of sports analytics and machine learning. Additionally we will explore paths for research. Moreover we will also delve into the implications of these findings within the realm of sports betting.

Chapter 2

Background and Literature Review

The purpose of this chapter is to explore the background and previous studies that form the basis of this thesis. It begins by examining the theories and concepts related to sports betting and predictive modeling. This not only establishes the context, for the thesis but also helps in understanding the scope of this study. The chapter then conducts a literature review analyzing the existing methodologies, their strengths, weaknesses and their effectiveness, in predicting football match outcomes well as implementing betting strategies. Additionally it assesses how machine learning techniques have been utilized in this field by presenting an analysis of research and their findings. This chapter serves as a foundation on which this thesis is built providing insights that inform subsequent experiments and discussions.

Football prediction and betting have been extensively researched, with an amount of literature examining facets of the issue. In this section we will provide a summary of the literature primarily concentrating on four key areas; predicting the outcome of football matches implementing effective betting strategies to find value utilizing algorithmic trading for in play betting and leveraging data from betting exchanges.

2.1. VALUE BETTING STRATEGIES

2.1.1. Football Match Outcome Prediction

Prediction of football match outcomes has been a subject of interest for many researchers. Traditional approaches have often relied on statistical techniques (Dixon and Coles 1997; Rue and Salvesen 2000). More recently, machine learning techniques have been applied to the problem, yielding promising results (Zhang et al. 2022; Fátima Rodriguesa 2022).

The Dixon and Coles model (Dixon and Coles 1997) is a recognized technique used to forecast the outcomes of football matches. This model utilizes a Poisson regression approach to consider the varying strengths of teams and the advantage of playing at home. It also addresses the problem commonly

encountered in Poisson models, where low scores are underestimated and high scores are overestimated. Although it may seem uncomplicated this model has demonstrated its efficacy by providing predictions for closely contested matches.

The double Poisson model is another method used to predict the outcomes of football matches. It builds upon the Poisson model by considering how the opponents defense can impact the number of goals scored by a team in a match (Karlis and Ntzoufras 2003). This model has proven to be more accurate, than the Poisson model in datasets.

On a note Rue and Salvesen (Karlis and Ntzoufras 2003) took an approach to tackle this problem. Their model takes into account factors such as team strength, home field advantage and the correlation, between the number of goals scored by both home and away teams. In scenarios this model has shown performance compared to the Dixon and Coles model especially when predicting matches where one team is clearly favored.

In times researchers have explored the application of deep learning methods to tackle this problem. Zhang et al. (Zhang et al. 2022) introduced an approach called the AS LSTM model, which combines the Long Short Term Memory (LSTM) model with an attention mechanism. They applied this model to predict football match results by utilizing a sliding time window and incorporating teams historical match data. This advanced approach demonstrates the potential of machine learning models, in forecasting sports outcomes.

Fatima et al. (Fátima Rodriguesa 2022) conducted an analysis of machine learning algorithms for predicting football match outcomes. They evaluated decision trees support vector machines, artificial neural networks and other models. The performance of these models varied depending on the features used and how they were configured.

As we transition from techniques to more sophisticated machine learning methodologies for predicting football match outcomes there is a growing need, for synthetic features that capture the nuances and dynamics of the games. These features play a role in unlocking the potential of machine learning models. A great example of this idea can be seen in how the Elo rating system's applied. Originally created for chess it has been adapted for events, like football. The Elo rating system is an used approach to determine the skill levels of players or teams in competition. It has proven effective in predicting match outcomes in football (Hvattum and Arntzen 2010). The system constantly updates a players rating based on game results with the magnitude of change depending on the difference, between players or teams ratings and the outcome of the game.

The Elo rating though powerful, on its own has served as the foundation for the development of intricate systems. One such example is the Pi ratings system (Constantinou and Fenton 2012) which's an adaptation of the Elo system. The Pi ratings system takes into account factors like match location (home, away or neutral) goal difference and recent match history. By considering these aspects the Pi ratings system creates a synthetic feature that encompasses various

aspects of a football match. This improved model has proven to be highly effective in predicting football match outcomes highlighting the potential of incorporating features to enhance prediction accuracy.

However despite these advancements there are still limitations and challenges when it comes to predicting football match results. One major challenge lies in the uncertainty, given the randomness present in football matches. With models at hand accurately predicting all match outcomes remains impossible due, to these unpredictable elements. Another challenge revolves around data quality and availability. The performance of these models heavily relies on quality and ample data used for training purposes. While an abundance of data exists for top tier leagues, lower tier leagues and non European leagues often lack reliable data.

More investigation is necessary to tackle these obstacles and delve into strategies, for predicting the outcomes of football matches. This might involve creating models that can effectively consider the unpredictability and randomness inherent, in football games exploring untapped sources of data and applying innovative machine learning methods.

2.1.2. Betting Strategies in Football

The development of betting strategies based on predictive models is another significant area of research. Value betting strategies, such as the one proposed by Szymanski (Szymanski and Kuypers 2003), have garnered considerable attention. The Kelly Criterion, a method for determining the optimal bet size, remains an enduring contribution to the field (Thorp 1966).

Szymanski and Kuypers (Szymanski and Kuypers 2003) introduced a value betting strategy centered around capitalizing on discrepancies between true match outcome probabilities and the odds provided by bookmakers. Their strategy dictates placing bets on outcomes that the market undervalues, aiming for a positive expected return. This methodology draws its foundation from the efficient market hypothesis, positing that consistently outperforming the average market return without incurring additional risk is implausible. Notwithstanding its intuitive nature, this tactic has empirically demonstrated efficacy in selected betting markets.

The Kelly Criterion, introduced by Thorp (Thorp 1966), is a well-known method for determining the optimal bet size. This criterion maximises the expected logarithm of wealth, which is equivalent to maximising the expected geometric growth rate of wealth. The Kelly Criterion is particularly effective when betting opportunities are independent and identically distributed, but it can be less effective in other situations.

Several variations and improvements have been proposed over the years to address the limitations and enhance the performance of the Kelly Criterion. One such variation is Fractional Kelly Betting. The Kelly Criterion can lead to large fluctuations in the size of the bettor's bankroll, which might not be tolerable to all investors due to their risk preferences. To manage this, some bettors use a fraction of the bet size suggested by the Kelly Criterion. This strategy, known as

Fractional Kelly Betting, can help manage risk but it also reduces the expected growth rate of the bankroll (MacLean, Ziemba, and Blazenko 1992).

More recently, Vlastakis, Dotsis, and Markellos (Vlastakis, Dotsis, and Markellos 2009) investigated the effectiveness of betting strategies in the football betting market. They found that simple betting strategies, such as betting on the home team or the favourite, are not profitable in the long run. However, more sophisticated strategies, such as those based on statistical models or machine learning algorithms, can yield positive returns.

Despite these advancements, there are still challenges in developing effective betting strategies. One of the main challenges is the inherent uncertainty in football matches, which makes it difficult to predict match outcomes with high accuracy. Another challenge is the bookmakers' margin, which reduces the expected return from betting. Moreover, the effectiveness of betting strategies can vary across different leagues and seasons.

Further research is needed to develop more sophisticated betting strategies that can overcome these challenges. This could involve the use of advanced machine learning techniques, the incorporation of in-play data, and the exploration of new betting markets.

2.2. MOMENTUM BETTING STRATEGIES

2.2.1. Use of Betting Exchange Data in Sports Betting

The utilization of betting exchange data in sports betting analytics has received substantial attention in recent research. Constantinou and Fenton (Constantinou and Fenton 2013) highlighted the utility of betting exchange prices as a potent predictor. They found that betting markets outperformed tipsters in predictive accuracy, emphasizing the immense value of betting exchange data as a source of insightful information for predictive modelling.

However, the application of machine learning techniques in sports betting, despite its promising results, comes with ethical and regulatory implications that warrant consideration. Vaughan Williams (Vaughan Williams 2014) provided an in-depth overview of sports betting regulation, and shed light on the potential for manipulation and fraudulent practices within betting markets. Similarly, Humphreys and Perez (Humphreys and Perez 2012) touched upon the ethical issues surrounding sports betting, including the risk of problem gambling and the potential impact on the integrity of sports. These works highlight the importance of mindful consideration of ethical and regulatory constraints when developing and applying machine learning models in sports betting.

In addition to the ethical and regulatory aspects of sports betting, there has been an increasing interest in the exploitation of betting exchange data to identify and capitalize on market inefficiencies. Clarke and Norman (Clarke and Norman 1995) delved into profitable betting strategies in football, focusing on utilizing betting exchange data to identify market biases. Their research indicated that the betting exchange market, while robust, still possessed inefficiencies which

could be exploited by bettors who possess comprehensive information. This pioneering work underscores the potential of betting exchange data not just as a reflection of collective sentiment, but also as a potent tool for predictive modelling and strategic application.

Moreover, beyond the predictive accuracy of betting markets, there has been significant research done on the dynamics of betting exchange markets themselves. For instance, Smith (Smith, Paton, and Vaughan Williams 2009) analysed the behavior of betting exchange prices over time and found that the market tends to overreact to new information, leading to potential profitable trading strategies.

2.2.2. In-Play Betting and Algorithmic Trading Techniques

In-play betting introduces a dynamic element to betting, where odds and opportunities change as the match progresses. Bunker and Thabtah (Bunker and Thabtah 2019) delved into the application of machine learning techniques for real-time prediction of sports outcomes, with a particular emphasis on football. Utilizing a combination of real-time statistics and past performance data, they constructed models that demonstrated potential in guiding in-play betting decisions, showing the power of data-driven methodologies in this dynamic betting environment.

Algorithmic trading, which involves the use of automated systems to place bets or trades, has also been applied to sports betting. Brown and Yang (Brown and Yang 2004) discuss the use of a betting exchange, akin to a stock exchange, to enable algorithmic trading in sports betting markets. Their study highlighted the similarities between financial markets and sports betting markets, and suggested that many of the techniques used in algorithmic trading could be applied to sports betting.

One of the key techniques in algorithmic trading is the use of time series analysis to predict future price movements. Chong, Han and Park (Chong, Han, and Park 2017) provided a comprehensive review of time series analysis techniques in algorithmic trading. Their exploration spanned traditional models such as ARIMA, state-space models, and transitioned into more advanced machine learning models, prominently including recurrent neural networks.

More recently, deep learning techniques have been applied to algorithmic trading. Dixon et al. (Dixon, Klabjan, and Bang 2017) used stacked LSTM networks, a type of recurrent neural network, to predict the mid-price movement of limit order books. Their results showed that the LSTM networks outperformed other machine learning models in terms of prediction accuracy and trading performance.

Despite these advancements, there are still significant challenges in the application of machine learning techniques to in-play betting and algorithmic trading. These include the inherent uncertainty and volatility of sports matches, the dynamic and complex nature of in-play betting markets, and the need for high-frequency, real-time data. Future research in this area could focus on addressing these challenges and further exploring the potential of machine

learning and deep learning techniques.

Chapter 3

Value Betting Strategy Framework in Football

This chapter dives deep into the first experiment of the thesis, shedding light on the process of synthetic feature creation from pre-match data, followed by the rigorous development of machine learning models tailored for predicting football match outcomes. The narrative unfolds the evolution of betting strategies, iterating from rudimentary fixed betting methods to the more intricate value betting approaches. The chapter meticulously outlines the adopted methodologies, presenting a transparent view into the research mechanisms. The results, are discussed in detail, offering insights into the predictive prowess of the developed models and the efficacy of the various betting strategies. Through this chapter, the reader gains a thorough understanding of the foundations of football betting, priming them for the subsequent chapters.

3.1. INTRODUCTION

In the ever-evolving landscape of sports analytics, the utilization of machine learning for predicting football match outcomes has emerged as a promising research frontier. The motive behind this exploration is twofold. Firstly, football, being the world's most popular sport, attracts significant betting interest. Improving the accuracy of match outcome predictions can potentially lead to enhanced betting strategies, thus providing an edge to punters in the competitive betting market. Secondly, the process of predicting football match outcomes presents an intricate problem due to the dynamic and complex nature of the sport. This complexity makes it an interesting and challenging problem for machine learning.

In the context of these motivations, this experiment aims to construct a comprehensive framework for implementing and refining machine learning models for football match prediction. We begin with the generation of a rich set of synthetic features based on pre-match data, which captures the multifaceted aspects influencing a match outcome. Following this, we design and implement various machine learning models, and evaluate their performance in predicting match outcomes.

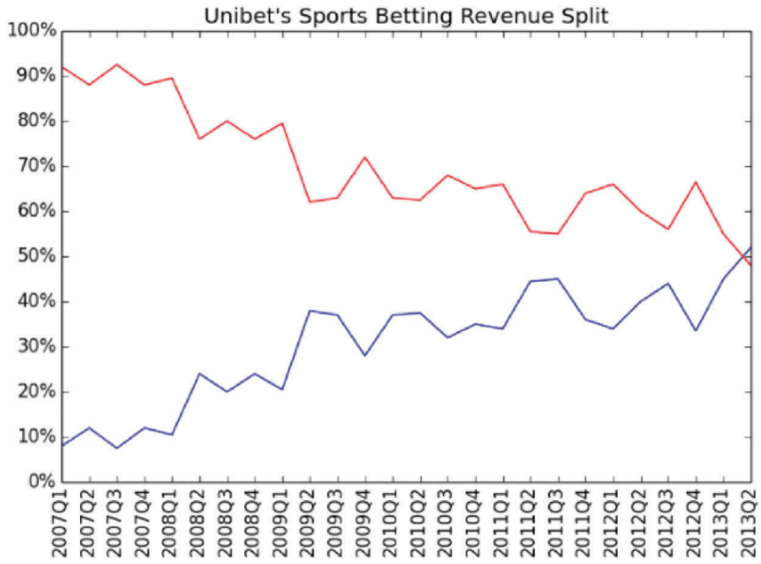

Figure 3.1: Unibet revenue distribution for football betting for In-Play and Pre-Game betting. (Divos et al. 2018)

The experiment also addresses the practical application of these predictions by devising a series of value betting strategies. Starting from simple fixed betting approaches, we iteratively enhance our strategies, incorporating more complex elements based on the predictions from our models. Through this, we demonstrate how machine learning predictions can be transformed into actionable betting strategies.

Overall, this experiment serves as a comprehensive exploration of the integration of machine learning models with value betting strategies in football, providing both theoretical insights and practical tools for football prediction and betting.

3.2. BACKGROUND

Football, often hailed as the beautiful game, has been intimately linked with betting for nearly as long as the sport itself has existed. From the informal wagers placed among fans in the late 19th century to the systematic football pools of the 1920s in the UK, the trajectory of football betting has mirrored the sport's meteoric rise in popularity. With the advent of televised matches in the latter half of the 20th century, the reach and scale of football betting underwent a significant transformation, embedding it deeply into the cultural fabric of the sport.

However, the essence of football lies in its inherent unpredictability. Beyond the visible parameters such as team form and player prowess, a myriad of external factors ranging from weather conditions to referee decisions can influence the trajectory of a match. This dynamism and unpredictability, while making the sport thrilling, posed significant challenges for those attempting to forecast match outcomes. Traditional betting methods, rooted in intuition and rudimentary statistics, often fell short in accurately capturing the multifaceted nature of football matches.

As computational capabilities expanded and data became increasingly accessible, a more systematic, data-driven approach to predicting match outcomes began to emerge. At the heart of this approach was the concept of value betting, which sought to capitalize on discrepancies between the odds offered by bookmakers and the actual probability of a given outcome. The real challenge lay in accurately determining this "true" probability.

The digital age brought with it a deluge of data. Everything from granular player statistics to broader team performances was now readily available, offering a detailed lens through which matches could be analyzed. This vast reservoir of data, while invaluable, also introduced complexities in terms of processing and analysis. Traditional statistical methods, while still relevant, often struggled to capture the nonlinear relationships and intricate patterns within the data.

This gap was bridged by the advent of machine learning in the realm of sports betting. Algorithms capable of handling vast datasets and discerning intricate patterns began to play a pivotal role. Machine learning models, ranging from linear regressions to more sophisticated neural networks, have revolutionized football predictions, offering enhanced accuracy and a deeper understanding of the game’s nuances.

In essence, the backdrop of this study is a rich tapestry of football history, the evolution of betting practices, and the transformative impact of data science and machine learning on the world of sports predictions. As we delve deeper into the subsequent sections, we will explore the modern methodologies and strategies that sit at this exciting intersection of sports, data, and technology.

3.3. DATA COLLECTION AND PREPROCESSING

3.3.1. Source of Data

The foundation of any data-driven research lies in its data source, and for this investigation, we turned to the esteemed "Football Data" website. Renowned in the realm of football analytics, this platform has consistently provided comprehensive datasets encompassing a multitude of leagues across various seasons. Its reliability, depth, and granularity make it an indispensable resource for researchers, analysts, and enthusiasts alike. This study specifically focuses in on data from three major football leagues: The English Premier League, the EFL Championship, and La Liga, aiming to capture a balanced mix of domestic and international matches.

3.3.2. Data Extraction

Given the nature of data spanning multiple seasons, careful attention was paid to ensure data consistency and integrity. The extracted datasets encompass several seasons, ensuring that the patterns and inferences drawn are robust, comprehensive, and not merely a reflection of a season-specific anomaly. It’s this breadth of data that provides a panoramic view of the evolving dynamics of football matches.

3.3.3. Features of Interest

A selective approach was adopted when choosing columns from the dataset. The aim was to ensure a judicious mix of features that offer insights both directly and indirectly into match outcomes. These columns capture various facets:

- Team information, match dates, and divisions to understand the context.
- Goal metrics to capture the essence of a match's outcome.
- Various match statistics that reflect team dynamics, aggression, and strategy.
- Betting odds from different bookmakers, providing a window into market sentiment and expectations.

Each metric, in its own right, offers a piece to the overarching puzzle of predicting match outcomes.

3.3.4. Synthetic Feature Generation

Synthetic features play a pivotal role in enhancing the informative capacity of a dataset. By deriving new features based on existing data, we can unearth deeper insights and potentially improve the predictive prowess of our models. In this study, several synthetic metrics were meticulously generated to capture the evolving dynamics of football matches. Some of these are outlined below.

Win Streak

A direct and intuitive metric, the win streak denotes consecutive wins by a team. It serves as an indicator of a team's current form and momentum. The formula is straightforward:

$$\text{Win Streak}_{\text{new}} = \begin{cases} \text{Win Streak}_{\text{old}} + 1 & \text{if team wins} \\ 0 & \text{otherwise} \end{cases}$$

This metric can potentially foreshadow a team's future performance, especially when used in conjunction with other metrics to capture a holistic view of a team's form.

ELO Ratings

Originally conceived in the realm of chess, ELO ratings provide a dynamic measure of a team's performance, taking into account both the result of a match and the relative strength of the opponents. The formula for updating a team's ELO rating after a match is given by:

$$ELO_{\text{new}} = ELO_{\text{old}} + K \times (S_{\text{actual}} - S_{\text{expected}})$$

where K is a constant factor, S_{actual} is the actual result of the match (1 for a win, 0.5 for a draw, 0 for a loss), and S_{expected} is the expected result based on the ELO

ratings of the two teams before the match.

The Pi-Ratings System

The Pi-Ratings system is an innovative football rating methodology, which borrows its foundational principles from the acclaimed Elo ratings introduced by Elo (Elo 1978). Primarily, this system aims to quantify the relative strengths of football teams. While retaining the foundational elements of the Elo system, the Pi-Ratings system incorporates modifications tailored to the unique dynamics of football.

Central to the Pi-Ratings system is a formula that dynamically updates after each match:

$$\Delta R = \alpha \left(W - \frac{R_A}{R_A + R_B^{\beta}} \right)$$

Where:

- R_A is the Pi-rating of team A (the home team) before the match.
- R_B is the Pi-rating of team B (the away team) before the match.
- W denotes the match result (1 for a win, 0.5 for a draw, and 0 for a loss).
- α is a constant determining the rate of rating change.
- β is a constant adjusting for the relative strengths of home and away teams.
- ΔR represents the change in team A's rating post-match.

The subsequent ratings after the match are derived as:

$$R'_A = R_A + \Delta R$$
$$R'_B = R_B - \Delta R$$

It's crucial to note that due to the addition of ΔR to team A's rating and its subtraction from team B's rating, the cumulative of all teams' ratings remains unchanged.

Advantages and Applications

The Pi-Ratings system introduces several enhancements over traditional Elo ratings:

- **Home Advantage:** Through the β parameter, the system accounts for the widely observed home advantage in football, where home teams typically outperform their away counterparts (R. Pollard and G. Pollard 2005).
- **Dynamic Ratings:** Ratings undergo updates post each match, rendering the system highly adaptive to fluctuations in team performances.
- **Scalability:** The system's design allows for its application across any football league or competition, accommodating a multitude of teams and matches.

The versatility of the Pi-Ratings system finds utility in diverse applications such as match predictions, team rankings, and even sports betting. By furnishing a relative strength metric for teams, it aids in driving more informed predictions and decisions in these domains.

GAP Ratings

The Generalised Attacking Performance (GAP) ratings system was developed to provide a more nuanced assessment of football teams' attacking and defensive capabilities. Unlike many traditional metrics, which may solely focus on the end results of matches, the GAP rating incorporates specific match statistics, such as goals scored, shots on target, and corners, among others.

The core of the GAP rating system is derived from its mathematical formulae, which consider both home and away performances. For a team indexed as i playing at home against a team indexed as j, the ratings are updated as follows:

$$
\begin{aligned}
H_i^a &= \max\left(H_i^a + \lambda\phi_1\left(S_h - \frac{H_i^a + A_j^d}{2}\right), 0\right) \\
A_i^a &= \max\left(A_i^a + \lambda\left(1-\phi_1\right)\left(S_h - \frac{H_i^a + A_j^d}{2}\right), 0\right) \\
H_i^d &= \max\left(H_i^d + \lambda\phi_1\left(S_a - \frac{A_j^a + H_i^d}{2}\right), 0\right) \\
A_i^d &= \max\left(A_i^d + \lambda\left(1-\phi_1\right)\left(S_a - \frac{A_j^a + H_i^d}{2}\right), 0\right)
\end{aligned}
$$

Similarly, for the opposing team j playing away against team i:

$$
\begin{aligned}
A_j^a &= \max\left(A_j^a + \lambda\phi_2\left(S_a - \frac{A_j^a + H_i^d}{2}\right), 0\right) \\
H_j^a &= \max\left(H_j^a + \lambda\left(1-\phi_2\right)\left(S_a - \frac{A_j^a + H_i^d}{2}\right), 0\right) \\
A_j^d &= \max\left(A_j^d + \lambda\phi_2\left(S_h - \frac{H_i^a + A_j^d}{2}\right), 0\right) \\
H_j^d &= \max\left(H_j^d + \lambda\left(1-\phi_2\right)\left(S_h - \frac{H_i^a + A_j^d}{2}\right), 0\right)
\end{aligned}
$$

In the formulae:

- H_i^a, A_i^a: Attacking ratings for team i for home and away games, respectively.
- H_i^d, A_i^d: Defensive ratings for team i for home and away games, respectively.
- S_h, S_a: Represents the attacking performance of the home and away teams.

- λ: A constant that dictates the impact of a single match on the ratings.
- ϕ_1, ϕ_2: Constants that balance the influence of a team's home performance on its away ratings and vice versa.

In practical applications, the GAP ratings system can be employed to rank teams, forecast match results, evaluate team performance, and guide decisions in the realms of sports management and betting. By offering insights into both attacking and defensive prowess, the GAP ratings afford a more comprehensive perspective on team performance compared to many traditional metrics.

3.4. MACHINE LEARNING MODEL SELECTION AND TRAINING

The goal is to not merely predict the outcomes of football matches. Instead, the focus lies in capturing the underlying probabilities of each possible outcome: a Home win, an Away win, or a Draw. This approach is pivotal in the context of sports betting. Here, the accuracy of a prediction is just one piece of the puzzle; equally important is the confidence with which this prediction is made. A model's ability to provide well-calibrated probabilities can significantly impact betting decisions, stakes, and ultimately, profitability.

3.4.1. Criteria for Model Selection

The nature of the problem at hand, predicting football match outcomes and their associated probabilities, demands a specific set of criteria for selecting the most suitable machine learning models:

- **Probabilistic Outputs**: Given the sports betting context, models that can provide probability estimates for each class (Home, Draw, Away) are preferred. These probabilities not only inform the predicted outcome but also guide betting stakes based on the confidence of the prediction.
- **Interpretability**: While complex models might offer higher accuracy, interpretability remains crucial. Being able to understand and explain model decisions can shed light on underlying patterns and strategies, aiding in trust and further refinement of the model.
- **Scalability & Efficiency**: Given the iterative nature of model training, especially with hyperparameter tuning, models that are computationally efficient and can scale with increasing data are preferred.
- **Robustness to Noise**: Sports data, especially football, can be noisy due to the myriad of unpredictable events during a match. The models should be robust enough to handle this noise and not overfit to specific anomalies.

3.4.2. Undestanding Market Efficiency

Implied odds, derived from a bookmaker's odds, offer a reflection of the perceived probability of a particular event or outcome. They can be calculated using the

formula:

$$P = \frac{1}{O}$$

Where P represents the implied probability and O is the offered odds. For example, if a bookmaker offers odds of 4.00 for a particular event, this translates to an implied probability of 25%. In an ideal betting market scenario, the summation of implied probabilities for all potential outcomes would be exactly 100%. However, in practice, bookmakers incorporate a margin known as the overround. This ensures that the total implied probabilities slightly exceed 100%, guaranteeing profit margins for the bookmakers irrespective of the outcome. The overround can be quantified using:

$$R = \sum_{i=1}^{n} \frac{1}{O_i} - 1$$

Where R represents the overround and O_i are the odds for each outcome. It's a measure of the bookmaker's profit margin and can shed light on the efficiency of the betting market. To get a true reflection of implied probabilities, one must adjust for this overround, ensuring that the probabilities sum up to 100% and better represent the actual likelihood of outcomes.

Significance in Predictive Modeling

Understanding market efficiency becomes critical when leveraging machine learning to predict match outcomes, especially when bookmaker odds serve as feature inputs. Several reasons underscore its significance:

1. **Benchmarking and Model Evaluation:** A highly efficient market implies that odds are a robust reflection of potential outcomes. In such a scenario, any model that merely mirrors this efficiency might just be capturing existing market knowledge without adding substantial predictive value.
2. **Influence on Feature Importance:** In an efficient market, odds are potent predictors, potentially overshadowing other features in terms of importance. Recognizing this dynamic can steer feature engineering efforts and model interpretability.
3. **Model Deviations and Insights:** A significant deviation of model predictions from efficient odds can be revealing. It might indicate potential overfitting or, more interestingly, the model's ability to identify and exploit market inefficiencies.
4. **Strategic Implications:** Market inefficiencies, when spotted, can be avenues for profit. Conversely, in markets that are near-perfect in their efficiency, the emphasis might not just be on predicting accurately, but also on effective risk management.

For each league we analysed the market efficiency using diverse analytical tools. Using La Liga League as an example, we employed various methodologies to examine how bookmakers' odds namely. Bet365 (B365), Betway (BW), Interwetten (IW) and Pinnacle Sports (PS), compared to actual game outcomes. Firstly, Frequency Matching was employed to evaluate how closely the

bookmakers' odds matched the actual frequency of match outcomes. The results were intriguing:

Bookmaker	Home	Draw	Away
B365	0.449	0.255	0.296
BW	0.447	0.255	0.297
IW	0.444	0.257	0.299
PS	0.450	0.254	0.295
Actual	0.457	0.260	0.283

Table 3.1: Frequency Matching of Bookmakers' Odds and Actual Outcomes

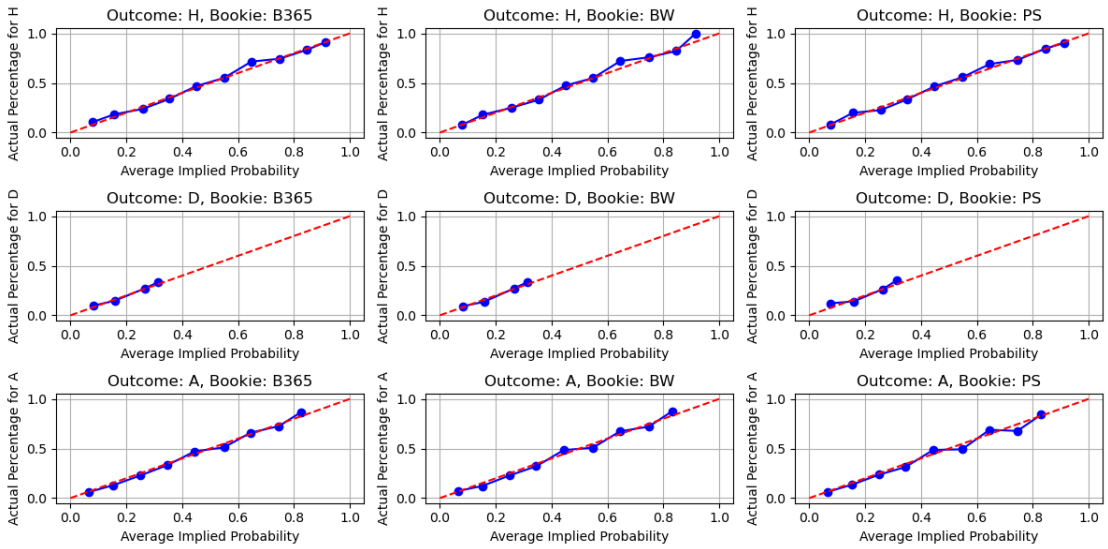

Figure 3.2: Calibration plot comparing predicted probabilities from bookmaker odds against actual match outcomes for La Liga games.

Our Calibration Plot visualized how the predicted probabilities from bookmakers aligned with the actual outcomes. Figure 3.2 provides a detailed depiction of this alignment, highlighting areas of over or underestimation for each outcome. The blue line represents the actual win percentage for each bin of implied probabilities. The red dashed line represents perfect calibration, where the implied probabilities would match the actual outcomes exactly.

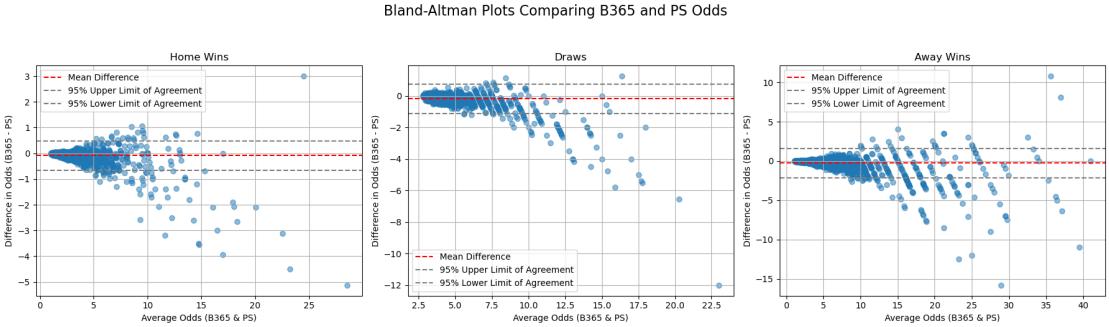

Figure 3.3: Bland-Altman plot showing the agreement between B365 and PS odds for home wins.

The Bland-Altman Plot was another tool used, specifically comparing the odds between bookmakers. This graphical method evaluates the agreement between two different ways of measurement. Figure 3.3 showcases the mean difference and the limits of agreement between Bet365 and Pinnacle Sports.

Lastly, the ROC AUC scores were assessed to understand the discriminative ability of the bookmaker odds in relation to actual outcomes. Values close to 1.0 indicate excellent discriminatory ability, while a value of 0.5 indicates no discrimination (i.e., no better than random guessing).

Outcome	B365	BW	PS
Home	0.721	0.721	0.721
Draw	0.584	0.587	0.590
Away	0.731	0.731	0.731

Table 3.2: ROC AUC Scores for Bookmakers' Odds (La Liga)

As shown above, for La Liga we are able to understand that while the bookmakers' odds demonstrate a good degree of market efficiency, as indicated by their proximity to actual frequencies and the above-average ROC AUC scores, they aren't perfectly efficient. There are minor discrepancies between the implied probabilities and actual frequencies, and disagreements between bookmakers as seen through the Bland-Altman Plot. Further, specific outcomes for given ranges of odds seem to be mispriced as shown through the calibrations plot. We repeated this analysis for each of the leagues and the findings are consistent, that while the market is reasonably efficient, there are nuances and potential inefficiencies that might be exploitable.

3.4.3. Data Splitting Strategy

The essence of any machine learning task lies not just in the model selection but also in how the data is treated and segmented. For this study, a chronological splitting of data was deemed most appropriate, considering the evolving nature of football, with teams' forms, strategies, and dynamics changing over seasons.

Dataset	Seasons	Leagues	No. of Matches
Training Data	2015-2021	EPL, EFLC & La Liga	1312
Validation Data	2021-2022	EPL, EFLC & La Liga	1312
Test Data	2022-2023	EPL, EFLC & La Liga	7872

Table 3.3: Data Splitting Strategy

3.4.4. Models Used in the Study

In this study, a chronological data segmentation strategy was employed to simulate real-world predictions, where past and present data guide predictions for future matches. A range of machine learning models known for their prowess in classification tasks was selected:

- **K-Nearest Neighbors (KNN):** A non-parametric method that leverages local patterns in the data by classifying based on the outcomes of similar matches.
- **Random Forest:** An ensemble method capturing diverse feature interactions, ideal given the multifaceted nature of football data.
- **Logistic Regression:** A statistical method providing insights into influential features, estimating outcome probabilities based on factors like team strength or home advantage.
- **Naive Bayes:** Efficiently computes outcome probabilities by assuming feature independence, suitable for extensive football datasets.
- **Boosting Algorithms (Gradient Boosting, XGBoost, CatBoost):** These algorithms iteratively correct past mistakes and excel in non-linear scenarios, making them pivotal for the complexity of football data.

3.4.5. Evaluation Metrics

Log loss, often termed as logarithmic loss, stands out as a pivotal metric when evaluating the performance of classification models, especially in contexts like sports betting where both the accuracy and confidence of predictions are paramount. Mathematically represented for binary classification as:

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)]$$

where y_i is the true label, p_i is the predicted probability, and N is the number of observations.

As discussed earlier the potential profitability of a bet isn't solely hinged on predicting the correct outcome but aligns closely with the model's estimated odds, which are reflective of perceived probabilities. Therefore, a model's capability to gauge the true underlying probabilities of outcomes is imperative. Log loss inherently promotes models that can do this accurately while cautioning against overconfidence, ensuring the development of models that are both accurate and reliably confident, aligning perfectly with the objectives of sports betting strategies.

3.5. BETTING STRATEGY DESIGN

Betting strategies in sports, particularly football, are crucial to manage and optimize the stake placed on individual bets. These strategies not only enhance the potential returns but also limit the risks associated with betting. As the complexity of the strategies evolves, they incorporate more information and provide finer control over the staking process, aiming to achieve a balance between potential gains and associated risks.

3.5.1. Fixed Fraction Strategy

This is one of the simplest strategies, often referred to as the "fixed wager" strategy. In the Fixed Fractional Betting approach, punters use a consistent stake size, usually a fixed percentage of their bankroll. This method reduces the risk of significant losses, especially during a series of unsuccessful bets.

$$\text{Stake} = \text{Bankroll} \times \text{Fixed Fraction}$$

The strategy's edge is further sharpened by focusing on Expected Value (EV). Instead of betting uniformly, wagers are placed on outcomes with the highest EV but only if this EV is greater than zero. This approach is grounded in the idea that over numerous bets, positive EVs will generally lead to profit, while negative ones lead to losses.

For outcomes $O_1, O_2, \ldots, O_n$ with expected values $EV_1, EV_2, \ldots, EV_n$:

$$\text{Chosen Outcome} = \max(EV_1, EV_2, \ldots, EV_n) \quad \text{if} \quad \max(EV_1, EV_2, \ldots, EV_n) > 0$$

In essence, the Fixed Fraction Strategy merges a systematic stake allocation with a value-driven betting approach, optimizing for both capital preservation and growth.

3.5.2. Kelly Criterion Strategy

The Kelly Criterion Strategy (Kelly 1956) is a betting and investment strategy that seeks to determine the optimal amount to wager on a given opportunity so as to maximize the long-term growth rate of the bankroll. The core philosophy of the Kelly Criterion is that if you can determine a definitive probabilistic edge over a betting market, then your stake should reflect this edge.

$$f^* = \frac{bp - q}{b}$$

Where:

- f^* is the fraction of the bankroll to wager
- b is the odds received on the bet decimalized
- p is the probability of winning
- q is the probability of losing (which is $1 - p$)

The result f^* is a fraction of the bankroll, and when applied, it can lead to optimal growth. However, if the estimated probabilities are off, it can also lead to significant losses.

3.5.3. Fractional Kelly Criterion Strategy

Recognizing the potential volatility and the substantial risk of ruin associated with the pure Kelly Criterion, the Fractional Kelly Criterion Strategy was introduced

(Thorp 1966). This strategy scales down the bet size recommendation of the Kelly Criterion by multiplying it with a fraction, usually less than 1.

The formula becomes:

$$\text{Stake} = \text{Fraction} \times f^* \times \text{Bankroll}$$

Where the fraction is chosen based on the bettor's risk appetite. A smaller fraction leads to a more conservative betting approach, while a larger fraction is more aggressive. Common fractions are half Kelly, quarter Kelly, etc.

3.5.4. Confidence Threshold Strategy

The Confidence Threshold Strategy is designed to place bets when the model's confidence in a prediction exceeds a certain threshold, and the predicted probability is greater than the odds-implied probability.

$$\text{Stake} = \begin{cases} 0 & \text{if Confidence} < \text{Threshold} \\ & \text{or Predicted Prob.} \leq \text{Odds-Implied Prob.} \\ \text{Bankroll} \times 0.01 & \text{otherwise} \end{cases}$$

3.5.5. Dynamic Fractional Kelly Criterion Strategy

Building on the Fractional Kelly, the Dynamic Fractional Kelly adjusts the fraction of the bankroll to bet based on the size of the bankroll and the perceived risk of the bet. The perceived risk is calculated as the absolute difference between the predicted probability and the odds-implied probability.

$$\text{Stake} = \begin{cases} \text{Lower Fraction} \times f^* \times \text{Bankroll} & \text{if Bankroll} < \text{Lower Threshold} \\ \text{Upper Fraction} \times f^* \times \text{Bankroll} & \text{if Bankroll} > \text{Upper Threshold} \\ \text{Standard Fraction} \times f^* \times \text{Bankroll} & \text{otherwise} \end{cases}$$

Additionally, the perceived risk of a bet is determined by the difference between the model's predicted probability and the odds-implied probability. The strategy may adjust the stake based on this perceived risk, betting less on riskier propositions and more on bets perceived as having value.

3.6. MODEL EVALUATION AND RESULTS

In our comprehensive endeavor to create predictive models for football match outcomes, datasets from three prominent football leagues — English Premier League (EPL), English Football League Championship (EFLC), and La Liga — were meticulously examined. Through the lens of each of these datasets, the generated synthetic features such as GAP ratings, ELO ratings, Pi ratings, and Win streak were scrutinized, forming the foundation for our subsequent modeling.

It is crucial to note that during model training, datasets from the three leagues were kept distinct and unmerged. We opted for this approach primarily because ratings across different leagues are intrinsically non-comparable. Thus, individual models were fashioned for each league to maintain their unique characteristics and dynamics.

Further augmenting our analytical approach, we incorporated the odds provided by various bookmakers. A comprehensive list of these bookmakers can be referenced below in table 3.4. Pregame odds from each bookmaker were used for model training as well as be

Bookmakers
Bet365
Interwetten
Betway
Pinnacle Sports

Table 3.4: List of Bookmakers

Following feature generation, several machine learning models as discussed earlier were trained on data from each league, with their performance being benchmarked using the log loss metric.

Model	EPL Log Loss	EFLC Log Loss	La Liga Log Loss
CatBoost	1.0754	1.1650	1.0944
XGBoost	1.1586	1.2111	1.1785
Gradient Boosting	0.9825	1.0535	1.0025
Random Forest	1.0149	1.0837	1.0255
Logistic Regression	0.9331	1.0438	0.9987
KNN	5.8123	5.3931	5.8828
Naive Bayes	2.6608	3.0510	2.6593

Table 3.5: Model Performance across leagues based on Log Loss

Having identified the best-performing models for each league, the focus pivoted towards assessing the profitability of our array of betting strategies. The strategies' performance was gauged across each league using the respective top models. To make betting decisions, we leaned on the English Premier League (EPL) as a representative example. From the previously described performance metrics, the three models that showcased the least log loss for EPL were Logistic Regression, Gradient Boosting, and Random Forest. This trifecta of models would then be the base for our further betting strategy simulations. Refer to the figure 3.4 below for a visual comparison of the log loss for each model:

To establish a baseline for the betting strategy, we utilized the simple Fixed Fractional Betting Strategy. This strategy, as mentioned earlier above, keeps the betting amount consistent in terms of a fraction of the current bankroll. Each strategy is given a starting bankroll of £1000. The detailed performance metrics for each model are illustrated in the following table:

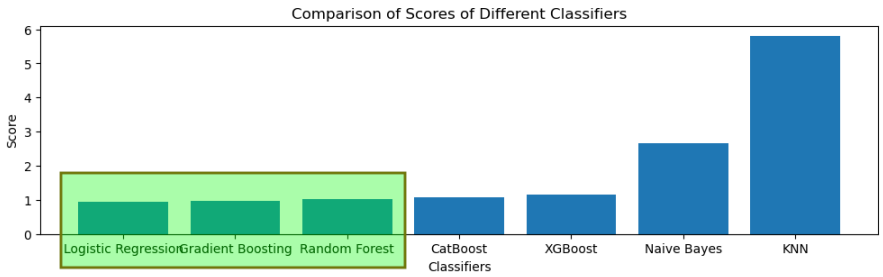

Figure 3.4: Selection of top 3 models based on Log Loss for EPL

Model	Profit	Bankroll	SD Returns	ROI	Yield	Win Rate	Volatility	Sharpe
LR	-265.85	734.14	0.0904	-0.2658	-0.0253	0.3324	47.178	0.0314
GB	39.48	1,039.48	0.0937	0.0394	0.0007	0.3693	318.556	0.0404
RF	-527.63	472.37	0.0976	-0.5276	-0.0126	0.3386	292.527	0.0204

Table 3.6: Performance metrics for the top 3 models using Fixed Fractional Strategy on EPL

For a holistic selection of the best model, one must consider a combination of these metrics, which are further explained in Section 3.7. Further, a visual assessment was also conducted, shedding light on the evolution of the bankroll & cumulative returns over time, distribution of returns, and win/loss stratified by bet type. This visualization can offer intuitive insights, aiding in decision-making and tweaking strategies as necessary.

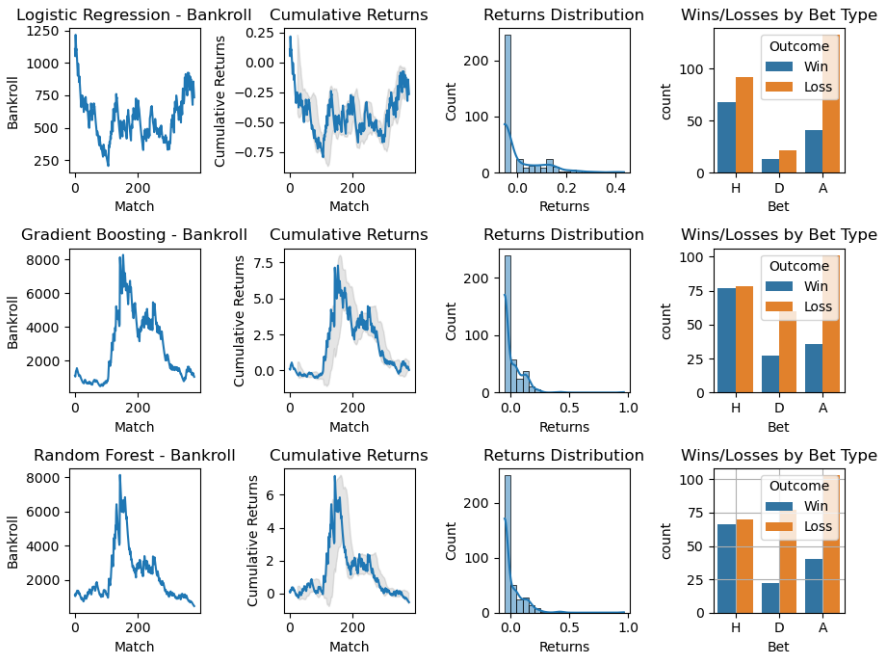

Figure 3.5: Visual representation of Bankroll, Cumulative Returns, and Win/Loss Distribution for top 3 models

Upon selection, the chosen model can then be integrated with more advanced strategies, expanding upon the initial Fixed Fractional Betting approach. These

advanced strategies, as mentioned earlier, can be catered to mitigate the risks further and capitalize on the winning streaks or adjust dynamically based on the perceived risks. This iterative process ensures the continuous refinement of the betting strategy, optimizing for better returns while managing potential downsides.

Given our analysis, it is apparent that the Gradient Boosting method was the most suitable model for the EPL dataset. Its selection was fortified by its positive ROI/Yield and superior win rate. Although it exhibited high volatility, the highest Sharpe ratio among the strategies presented a favorable risk-adjusted return, warranting its preference.

Building upon the foundation of the Fixed Fractional Strategy, we expanded our strategic arsenal by iterating more advanced strategies, all applied on the Gradient Boosting model. The results for each of these strategies, viz. Confidence Threshold Strategy (CT), Kelly Criterion Strategy (K), Fractional Kelly Criterion Strategy (FKC), and Dynamic Fractional Kelly Criterion Strategy (DFKC), are concisely tabled below:

Strat.	Profit	Bankroll	SD Returns	ROI	Yield	Win Rate	Volatility	Sharpe
CT	-430.48	569.52	0.0873	-0.431	-0.034	0.5935	73.896	0.0021
KC	-840.60	159.40	0.2182	-0.841	-0.022	0.3694	238.00	0.0744
FKC	1069.50	2069.50	0.0327	1.07	0.095	0.3694	52.939	0.0744
DFKC	2014.53	3014.53	0.0701	2.015	0.053	0.3694	189.66	0.0742

Table 3.7: Performance metrics for advanced strategies using Gradient Boosting model on EPL

Again from a visual perspective, it's beneficial to understand how the bankroll evolved, how returns fluctuated, and how win/loss patterns emerged for each strategy. This approach assists in finding the most suitable and tested strategy

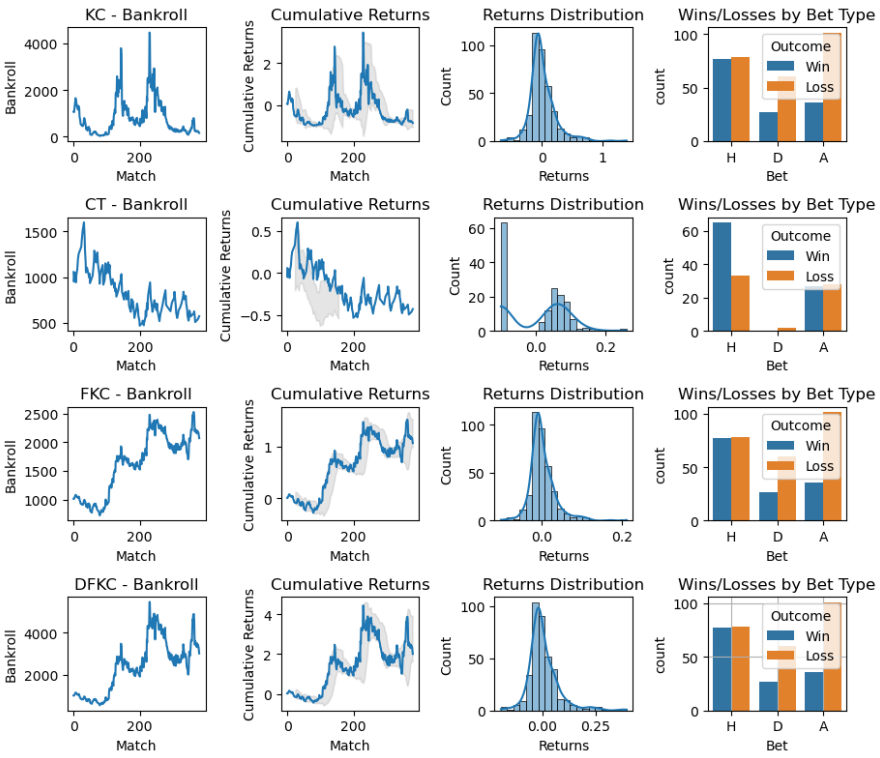

Figure 3.6: Visual representation of Bankroll Evolution, Returns, and Win/Loss Patterns for each advanced strategy

Having established this comprehensive approach for the EPL dataset, it is imperative to replicate this for the other leagues. A similar methodology ensures consistent treatment across datasets, allowing for valid cross-league comparisons. For each league:

- **Model Selection**: Pinpoint the top-performing model by considering metrics like ROI, Yield, and Sharpe Ratio.
- **Advanced Strategy Iteration**: Implement the roster of advanced strategies on the selected model, deriving performance metrics for each strategy.
- **Visual Analysis**: Furnish visual depictions of bankroll evolutions, return distributions, and win/loss patterns to provide an intuitive understanding of each strategy's performance.
- **Conclusion and Insights**: Based on gathered data, draw league-specific conclusions. This step might also entail tweaking strategies to better suit league dynamics or incorporating league-specific features into the model.

By maintaining a consistent methodological approach across all leagues, we can effectively assess the versatility and robustness of our chosen models and strategies. This rigorous approach aids in mitigating overfitting to a particular dataset and ensures a more generalizable betting strategy that holds promise across multiple football leagues. Below are the metrics and the visual analysis of the best performing models

3.6.1. Results of model on English Football League Championship (EFLC)

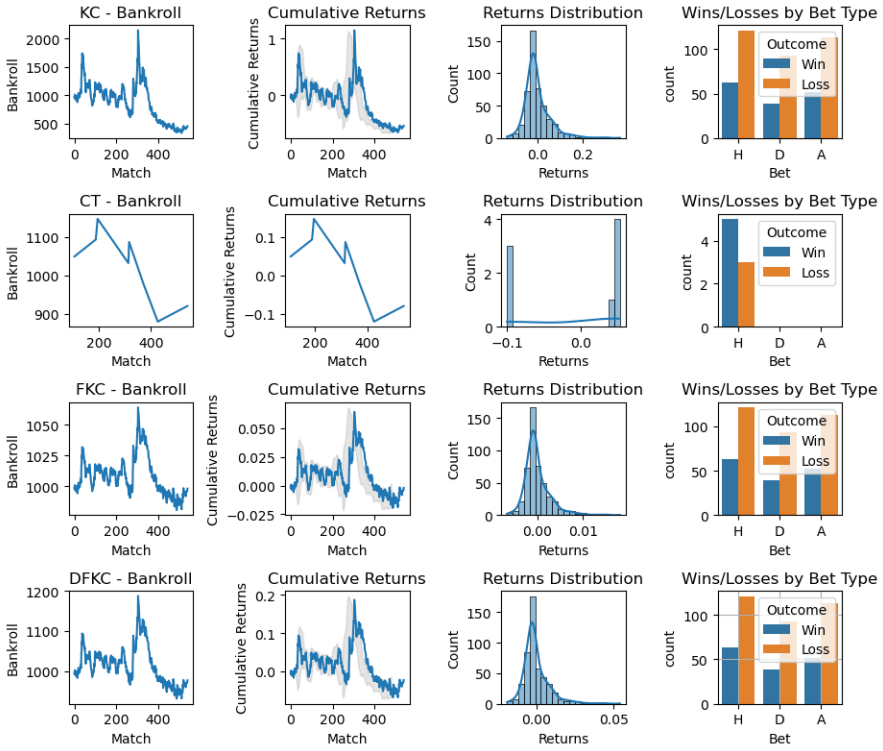

Figure 3.7: Visual representation of Bankroll Evolution, Returns, and Win/Loss Patterns for each advanced strategy on EFLC

Model	Profit	Bankroll	SD Returns	ROI	Yield	Win Rate	Volatility	Sharpe
Base	-399.61	600.39	0.0282	-0.40	-0.043	0.3477	25.294	-0.0201
KC	-545.47	454.53	0.0593	-0.546	-0.041	0.3202	59.057	0.00003
CT	-79.33	920.67	0.0766	-0.079	-0.096	0.6250	80.689	-0.0996
FKC	-2.05	997.95	0.0030	-0.002	-0.003	0.3202	2.993	0.00003
DFKC	-22.76	977.24	0.0085	-0.023	-0.011	0.3202	8.754	-0.0014

Table 3.8: Performance metrics for advanced strategies using fine-tuned gradient boosting model on EFLC

For the EFL Championship (EFLC) dataset, a detailed analysis of the results offers some noteworthy observations. Despite rigorous model fine-tuning and strategic adjustments, the selected betting strategies have not generated profitable outcomes, however have only minimised losses. This contrasts sharply with the results observed in the English Premier League (EPL), where strategies showcased better performance. One potential inference from these findings is that the dynamics and unpredictability inherent in the EFLC games might demand a more nuanced approach. To achieve better predictive accuracy, further feature engineering is recommended. It's also worth considering the exploration of alternative machine learning models and possibly develop a more

proficient model that can accurately predict expected match outcomes in the EFL Championship.

3.6.2. Results of model on La Liga League

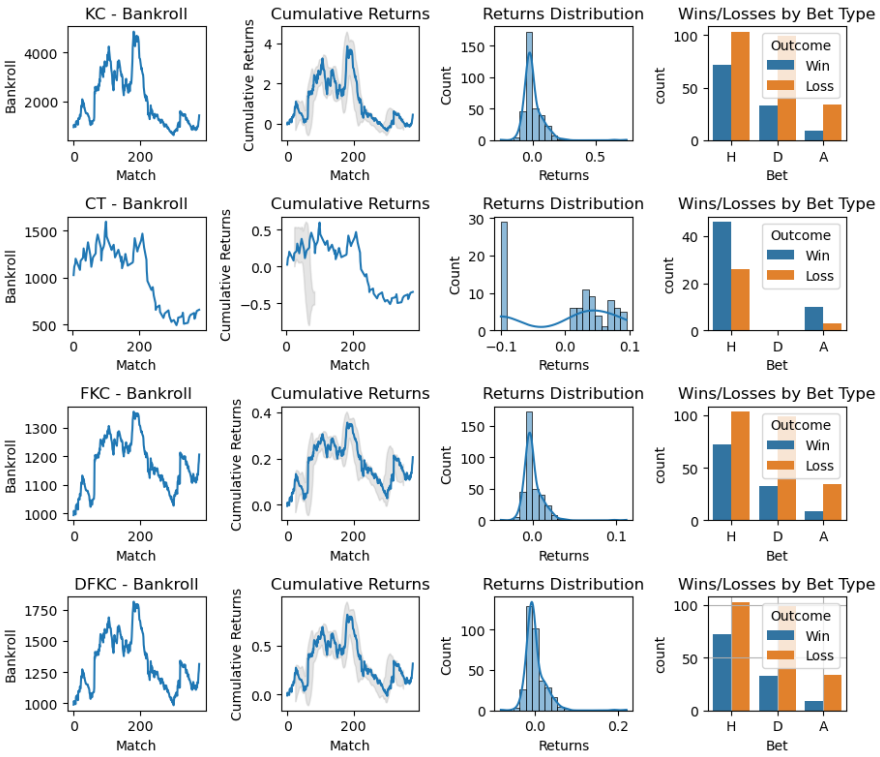

Figure 3.8: Visual representation of Bankroll Evolution, Returns, and Win/Loss Patterns for each advanced strategy on La Liga

Model	Profit	Bankroll	SD Returns	ROI	Yield	Win Rate	Volatility	Sharpe
Base	-174.00	826.00	0.0328	-0.174	-0.023	0.3257	37.269	-0.0008
CT	-343.29	656.71	0.0738	-0.343	-0.040	0.6588	81.433	-0.0295
KC	439.09	1439.09	0.0843	0.439	0.014	0.3257	151.38	0.0485
FKC	205.99	1206.00	0.0126	0.206	0.072	0.3257	14.244	0.0485
DFKC	314.48	1314.48	0.0258	0.314	0.047	0.3257	32.297	0.0425

Table 3.9: Performance metrics for advanced strategies using Logistic Regression model on La Liga

For the La Liga dataset, the Logistic Regression model was chosen due to its ability to capture the intricacies of match outcomes within the league and its performance compared to the other models for the baseline fixed fractional betting strategy. Each football league has its own dynamics influenced by various factors. The Logistic Regression model has shown its proficiency in analyzing La Liga's unique blend of these factors, thereby proving to be the most suitable model for this dataset. The subsequent application of advanced betting strategies allowed us to optimize the potential returns, with each strategy

yielding its own set of outcomes as presented in the table.

3.7. ANALYSIS AND DISCUSSION

The results discussed in the preceding section lay the foundation for a comprehensive discussion on the predictive capability of our models, their application in betting strategies, and the inherent differences across the football leagues under study. This section dives deep into the analytics, drawing insights and discussing implications.

3.7.1. Comparative Performance of Predictive Models

A variety of machine learning models were deployed on our datasets, from which notable disparities in predictive performance were evident. The low log loss values for models such as Logistic Regression, Gradient Boosting, and Random Forest, especially when applied to the English Premier League dataset, highlight their robustness in prediction. However, models like KNN and Naive Bayes showed significantly higher log loss, indicating poorer prediction accuracy for football match outcomes. This may underscore the significance of feature interactions and non-linearities that the latter models are not aptly capturing.

In juxtaposition, the role of bookmakers' odds cannot be undermined. By integrating pregame odds into model training, we have leveraged an external benchmark that harnesses the collective wisdom of the masses. This integration likely improves the predictive performance, as the odds encapsulate a myriad of factors, including public sentiment, recent team performance, and expert analysis.

3.7.2. Strategic Betting: An EPL Perspective

Our results reflect the symbiotic relationship between predictive modeling and its application in strategic betting. While the Logistic Regression, Gradient Boosting, and Random Forest models exhibited superior performance on the EPL dataset, it was Gradient Boosting that ultimately stood out when the betting lens was applied.

Its positive ROI/Yield and superior win rate, despite high volatility, suggest that while betting based on this model's predictions can be a roller-coaster ride, it tends to be more profitable in the long run. The Sharpe ratio further reinforces this, indicating that the risk-adjusted return is favorable.

Moreover, the shift from the baseline Fixed Fractional Betting Strategy to more advanced strategies elucidates the importance of dynamic decision-making in betting. The Dynamic Fractional Kelly Criterion Strategy (DFKC) performed exceptionally well, turning an initial bankroll of £1000 into a staggering £3014.53. This strategy's adaptability — adjusting bet amounts based on perceived risks and potential returns — might be its silver bullet.

3.7.3. Inter-League Variations

The comparison between the English Premier League and the English Football League Championship offers intriguing insights. The former, being a more popular league, might benefit from a wealth of data, expert analysis, and more predictable gameplay. On the contrary, the EFL Championship's dynamics and inherent unpredictability might contribute to the challenging predictive landscape, as evidenced by our results.

Furthermore, the La Liga results (which were truncated in the previous section) would undoubtedly have added another dimension to our inter-league analysis. Different cultural, strategic, and gameplay nuances might render one predictive model more effective for one league over another. It reinforces the importance of tailoring models and strategies to the unique characteristics of each league, rather than adopting a one-size-fits-all approach.

3.7.4. Refinement and Forward Path

The results achieved provide a compelling case for the integration of machine learning in sports analytics and betting. Yet, there's ample scope for refinement. The EFLC results, for instance, warrant deeper exploration into feature engineering, alternative machine learning models, and more bespoke betting strategies.

Furthermore, additional data sources, such as player statistics, weather conditions, or even historical head-to-head matchups, can be incorporated to enhance the predictive prowess of our models.

3.7.5. Conclusion

In summary, our journey through predictive modeling for football outcomes underscores the potential of machine learning in sports analytics and betting. With rigorous validation, tailored strategies, and constant refinement, such models can be both a fascinating academic endeavor and a profitable venture.

However, it's crucial to acknowledge the uncertainties inherent in sports. No model, no matter how sophisticated, can guarantee consistent profits. Responsible betting, informed decision-making, and a constant thirst for learning and adaptation should guide our future endeavors in this domain.

3.8. SUMMARY

In this experiment, we embarked on a journey to understand the multifaceted dynamics of football matches and unravel how specific features influence the outcomes of these matches. The overarching aim was not merely predictive, but to derive valuable betting strategies that can capitalize on the predicted outcomes. This is particularly crucial given the immense betting interest in football, the world's most popular sport.

Our approach was informed by rigorous theories from existing literature, analyzing

the relationships between various team performance metrics and the results of matches. Unique to our study, we extended beyond just core outcome-focused metrics, encompassing a wide range of influencing factors. This was done to ensure an overall understanding, rather than a myopic view of match outcomes. Additionally, our exploration touched on pivotal concepts in the betting world, such as the Kelly Criterion, emphasizing its applicability and value in deriving betting strategies.

Moreover, one of the cornerstones of our experiment is the development of an extendable framework. This framework is designed to facilitate the ease of training and testing models and strategies specifically tailored for pregame football betting. Thus, beyond the immediate insights and strategies derived from our study, we also offer a robust tool for future explorations in this domain.

In conclusion, this experiment serves as an effort in bridging the worlds of machine learning and football betting. While our findings provide immediate strategies and insights, the true value lies in the holistic approach and the extendable framework we've developed, promising a robust foundation for future endeavors in football analytics and betting strategies.

Chapter 4

Momentum Betting and Drift Betting

Venturing into the dynamic world of in-play betting, this chapter discusses the second experiment of the thesis. It encapsulates the intricate process of harnessing betting exchange data to anticipate in-play events and craft momentum betting strategies. The emphasis shifts to the advanced application of stacked LSTM networks, drawing inspiration from the realm of algorithmic trading. The methodological journey is charted with precision, offering readers a comprehensive insight into the research's underpinnings. Results, shaped by the intersection of data science and sports analytics, are dissected, revealing the ability of the in-play predictive models and the viability of the momentum betting paradigms. This chapter not only expands the reader's understanding of in-play betting intricacies but also sets the stage for the culmination of the thesis, intertwining the learnings from both experimental explorations.

4.1. INTRODUCTION

In the vast domain of sports betting, the real-time, in-play betting arena represents an exciting and challenging frontier. This dynamic mode of betting, where wagers are placed while the event is ongoing, necessitates swift and informed decision-making. Two primary motivations drive this experiment. Firstly, the unpredictable nature of live football matches, with their constantly shifting momentum, offers a solid basis for predictive modelling. Accurate real-time predictions can aid in devising swift betting strategies, providing punters an advantage in the rapid-paced in-play betting environment. Secondly, the challenge lies in processing live data streams efficiently and making predictions under time constraints, making it a compelling problem for advanced machine learning models like stacked LSTMs.

Within this context, our experiment endeavors to harness the power of in-play data, both collected through our custom-built listener and procured from premium sources like Betfair Exchange. We aim to sculpt this data into predictive models, with a special emphasis on stacked LSTM networks, renowned for their prowess in time-series forecasting.

The practical application of our real-time predictions unfolds as we construct a robust backtesting framework. This tool allows us to iteratively refine our betting strategies and run detailed simulations, thus mimicking real-world trades on football matches. By leveraging only Betfair in-play odds data, our stacked LSTM model endeavors to predict odds shifts, revealing potential profitable betting windows.

In essence, this experiment delves deep into the world of in-play betting, blending cutting-edge machine learning techniques with momentum and drift betting strategies. The journey provides both a theoretical exploration of real-time football match prediction and hands-on strategies for in-play betting, bridging the gap between advanced analytics and real-world application.

4.2. BACKGROUND

Football's dynamic nature is not only encapsulated in the 90 minutes leading to the final whistle but is significantly amplified during the live action of the game. The suspense of real-time events – a pivotal goal, a red card, or even a penalty kick – injects an adrenaline rush both on the field and among the fans. This real-time thrill has given rise to the thriving domain of in-play betting, where punters place bets as the action unfolds on the pitch.

Historically, betting was confined to predictions made before the commencement of a match. However, the evolution of technology and live broadcasting reshaped this paradigm, allowing bettors to engage with matches in real-time. The allure of in-play betting lies in its immediacy and the plethora of opportunities it presents within a single match. With advanced models such as gradient boosted trees showing significant promise in sports analytics (Hubáček, Šourek, and Železný 2019), the foundation for incorporating machine learning and deep learning into in-play betting was laid. A game's momentum can shift rapidly, and with it, the betting odds fluctuate, offering keen punters numerous windows of opportunity.

Yet, the challenge of in-play betting is monumental. The rapid pace at which events unfold requires not only swift decision-making but also a deep understanding of the game's current dynamics. Traditional methods of intuition-based betting often falter in this high-octane environment, as they cannot adapt quickly to the game's shifting sands.

Enter the world of algorithmic trading and machine learning. Inspired by the stock market's high-frequency trading systems (Aldridge 2013), which leverage algorithms to make rapid-fire decisions, the sports betting domain witnessed the integration of similar technologies. Using live data streams and sophisticated predictive models, like the stacked LSTMs, the task of predicting in-play events and corresponding betting opportunities became more systematic and data-driven.

The crux of this experiment's background is rooted in the evolution of in-play betting, the technological advancements that facilitated its rise, and the challenges and opportunities it presents. As we traverse this chapter, we'll navigate the intricate maze of momentum and drift betting, harnessing the power

of deep learning models and real-time data. Embracing methodologies inspired by algorithmic trading and cutting-edge data analysis, this chapter aims to explore the uncharted territories of in-play football betting at the intersection of technology, data, and sports.

4.3. DATA COLLECTION AND PREPROCESSING

4.3.1. Historical Data from Betfair Premium

Betfair Premium provides a renowned service offering a comprehensive repository of historical data on various sports markets. For our research's intent and purposes, we are leveraging the rich details provided by Betfair Premium related to the match outcome markets.

The dataset, which was sourced from Betfair Premium, contains data from May 2020 and January 2023, encapsulating an extensive array of matches and the associated betting information. This data predominantly covers the odds markets for football match outcomes, ensuring a perfect alignment with our research objectives. This dataset not only provides the odds for each possible outcome (Home, Away, Draw) but also details the betting volumes for each runner.

4.3.2. Listener for Live Market Data Collection

To complement the historical data procured from Betfair Premium and to capture the dynamic essence of in-play betting, we designed a custom listener. This listener's primary objective is to record real-time market data, capturing the evolving odds and volumes as the matches unfold.

The listener has been meticulously designed to interface seamlessly with Betfair's live data streams. Its core functionality revolves around tapping into the live data streams of ongoing matches and recording pertinent data points in real-time. Figure 4.6 shows the design of the market listener and it's ability to record multiple markets using multiple threads to handle different streams. Further, the listener also has the ability to store recorded games to database or local file and is written to be extensible to other storage methods.

Figure 4.1: Design of Betfair Market Listener

The output from the listener is not only restricted to live odds for each match outcome but also encompasses the betting volumes associated with each runner

and stored in the same format as the historical data. The continuous operation of the listener during matches, capturing data at regular intervals, ensures a comprehensive and detailed dataset, mirroring the volatile nature of in-play betting.

An important component of our listener system is the data visualizer, built using Streamlit. This tool offers an intuitive interface to visualize the live-streamed data post-acquisition. Not only does it allow users to gain immediate insights into the data's structure and patterns, but it also facilitates data verification. Recognizing the importance of data integrity in predictive modeling, this visualizer comes equipped with a data quality check feature. Users can generate comprehensive reports that assess the quality of the acquired data, ensuring its reliability and readiness for further analysis. This seamless integration of visualization and quality verification makes the tool indispensable for our data acquisition and preprocessing workflow.

Figure 4.2: Screenshot of the Streamlit-based Data Visualizer showcasing the live-streamed football betting data

4.3.3. Odds Dynamics and In-Game Events

During live football matches, odds exhibit significant volatility in response to in-game events, offering a dynamic representation of the shifting probabilities of match outcomes. Taking the Crystal Palace v Brighton game on 11-02-23 as an illustration (figure 4.3), the odds fluctuations provide a vivid narrative of the match's progression. When Brighton netted the first goal, a pronounced spike in the odds for a Crystal Palace win reflected the diminished probability of that outcome. Conversely, as Crystal Palace equalized later in the match, the odds saw a sharp contraction, echoing the renewed hope for a Crystal Palace victory. This ebb and flow of odds not only capture the real-time pulse of the game but also offer a surrogate marker for significant events, such as goals, highlighting the intertwined relationship between on-pitch actions and market perceptions.

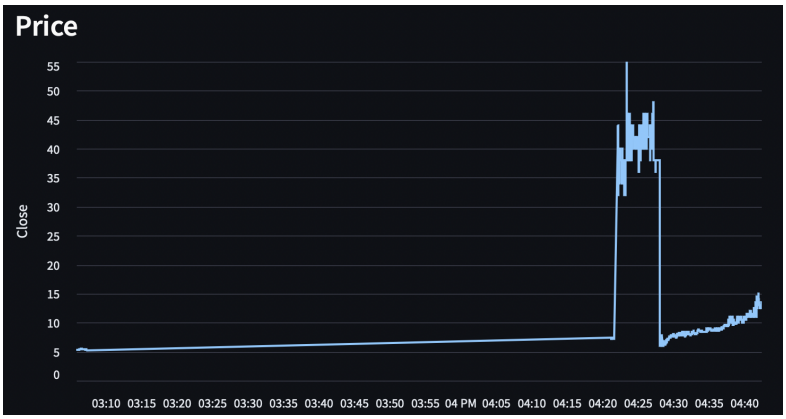

Figure 4.3: Odds Volatility for Crystal Palace during the Crystal Palace v Brighton Match on 11-02-23

The volatility in odds during a live match introduces an inherent noise into the data. This noise is a byproduct of various unpredictable elements within the game, such as unexpected player behaviors, referee decisions, or even crowd reactions. While significant events like goals can cause discernible shifts in the odds, the continuous oscillations in between arise from the market's collective attempt to forecast the unpredictable nature of the game in real-time. Such noise poses challenges for predictive modeling, as it can mask true underlying patterns and reduce the clarity of signals essential for accurate predictions.

4.3.4. Data Pre-Processing

Data pre-processing forms a crucial part of any machine learning endeavor. It involves refining the raw data into a format that can be readily consumed by our LSTM model, ensuring the subsequent model accuracy and reliability.

- **Data Cleaning:** Our initial step post-collection was to thoroughly clean the data.
  - *Handling Missing Values:* Despite the robustness of our data sources, we identified sporadic missing values. Such gaps were systematically addressed using interpolation methods, ensuring the data continuity.
  - *Outlier Detection and Rectification:* An initial exploratory data analysis helped identify potential outliers or anomalies, especially in the live data. Contextual understanding was employed to rectify these outliers, ensuring a more representative dataset.
- **Feature Engineering:** With the foundational cleaning in place, we delved into feature crafting.
  - *Synthetic Features:* While our primary features were sourced directly, certain synthetic features were considered, aiming to encapsulate latent patterns or relationships.
- **Normalization:** Given the diverse range of odds and volumes in our dataset, normalization was essential.

  - *Min-Max Scaling:* We employed Min-Max scaling to bring all features onto a similar scale. This process ensures equitable consideration of all features during the modeling phase.

- **Data Integration:** The final step in our pre-processing journey was data merging.
  - *Merging Data Sources:* Data from Betfair Premium and our custom listener were seamlessly integrated. This integration involved aligning data points based on unique timestamps and match identifiers.
  - *Consolidated Dataset:* Post-integration, our dataset housed information from approximately 6093 matches, serving as the foundation for subsequent model training.

4.3.5. Synthetic Features

Synthetic features are derived features that encapsulate specific trends, patterns, or relationships in the data. In the context of this study, several synthetic features were engineered to capture various dynamics in the betting odds data. The following are the primary synthetic features constructed:

1. **Rate of Change (RoC):** The Rate of Change is a metric that evaluates the percentage change in data over a specified period. It provides insights into the momentum and the strength of the data's movement. Mathematically, for a given period p, it is defined as:

$$\text{RoC} = \frac{\text{Value at time } t - \text{Value at time } t-p}{\text{Value at time } t-p}$$

2. **Moving Average (MA):** The Moving Average smoothens the data by taking an average of the data points over a specified window, thus highlighting trends over short-term fluctuations. For a window of size w, the moving average is given by:

$$\text{MA}(t) = \frac{\text{Sum of values from time } t-w+1 \text{ to } t}{w}$$

3. **Moving Average Convergence Divergence (MACD):** MACD is a trend-following momentum indicator that shows the relationship between two moving averages of a dataset. It consists of the MACD line, which is the difference between short and long Exponential Moving Averages (EMAs), and a signal line. The MACD line and the signal line are defined as:

$$\text{MACD line} = \text{Short EMA} - \text{Long EMA}$$

$$\text{Signal line} = \text{Moving Average of MACD line over signal window}$$

   The MACD is particularly useful in identifying potential buy or sell signals when the MACD line crosses above or below the signal line.

These synthetic features, crafted from the raw odds data, play a pivotal role in enhancing the predictive prowess of our machine learning models by capturing intricate relationships and trends in the data.

4.3.6. Betting Framework

The integration of a robust betting strategy framework is pivotal for evaluating the efficacy of any predictive model in a real-world scenario. To this end, the extension of our framework to encompass backtesting and simulation capabilities using the ***flumine*** and ***betfairlightweight*** packages stands as a significant enhancement. Backtesting, a cornerstone of any trading strategy, allows us to simulate bets on past games, providing a historical perspective on the model's potential profitability. Through these simulations, also known as paper trades, we can gauge the model's performance without the financial risks associated with real betting. This iterative testing offers invaluable insights, enabling us to refine our strategies based on historical outcomes. Figure 4.4 demonstrates the use of the back testing framework with a trading strategy that uses the exponential moving averages as a signal for a specific game.

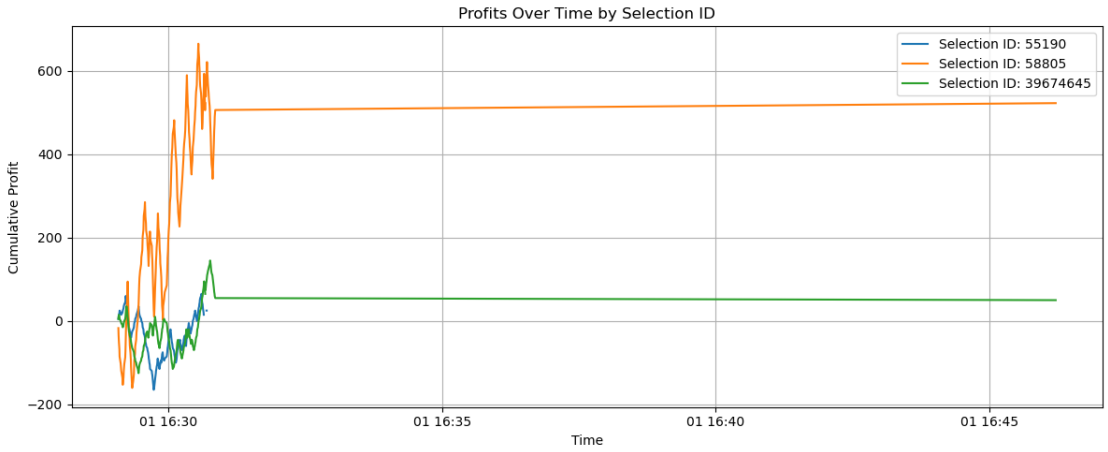

Figure 4.4: Cumulative Profits by Selection ID Over Time, using the Back Testing Framework

Furthermore, the augmented framework isn't limited to retrospective analysis. It offers the flexibility to trade on live games, simulating in real-time how the model would perform under current match conditions. This live simulation paints a holistic picture, combining historical insights with real-time events to provide a comprehensive view of the betting strategy's potential. The added ability to log profits of strategies ensures continuous tracking, allowing for timely interventions and strategy recalibrations. In essence, this extended framework bridges the gap between theoretical modeling and practical application, ensuring that our betting strategies are not just mathematically sound but also practically viable.

4.4. DESIGN OF GRM & LSTM ARCHITETURES

Predicting football match outcomes is inherently a sequential problem, where past events can provide insights into future outcomes. Deep learning models,

particularly Recurrent Neural Networks (RNNs), have shown great potential in handling sequential data. Among the various RNN architectures, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) stand out due to their ability to capture long-term dependencies in data. In the context of football betting, where the aim is to predict odds for three distinct outcomes (Home, Away, Draw), these architectures are especially relevant.

4.4.1. LSTM Model

LSTM, introduced by Hochreiter and Schmidhuber (1997), is designed to overcome the vanishing gradient problem of traditional RNNs. It achieves this through its unique cell state and three gates: input, forget, and output.

In our context, we employ a *stacked* LSTM model, where multiple LSTM layers are stacked on top of each other. This approach increases the depth of the model, allowing it to capture more complex temporal relationships. Stacked LSTMs have been traditionally employed in financial markets to forecast stock prices, owing to their ability to recognize intricate patterns across varying time scales.

The stacked LSTM processes a sequence of past odds and momentum related features to predict the odds for the three outcomes. By retaining memory from previous matches and odds, it can discern patterns and trends that can influence future odds. Stacked LSTMs are particularly powerful due to their hierarchical structure. Each layer can potentially learn different temporal representations, with lower layers capturing short-term dependencies and higher layers capturing long-term trends.

- **Input Dimension:** Represents the number of features in the input, such as past odds, team statistics, and player metrics.
- **Hidden Dimension:** Reflects the model's capacity, determining the number of LSTM units.
- **Output Dimension:** Set to 3, corresponding to the odds for Home, Away, and Draw outcomes.

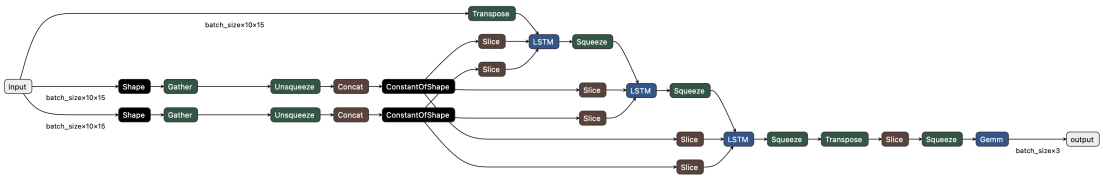

Figure 4.5: LSTM Model Architecture

4.4.2. GRU Model

GRU, proposed by Cho et al. (2014), is a variation of LSTM and is designed to be computationally more efficient by reducing the complexity of the LSTM unit. It combines the forget and input gates into a single "update gate".

For football betting, the GRU model also processes a sequence of past odds and match features to predict the outcomes. Its design allows it to capture essential

patterns in the data while being faster and requiring fewer parameters than LSTM.

- **Input Dimension:** Same as the LSTM model, it encompasses features like past odds and team metrics.
- **Hidden Dimension:** Represents the number of GRU units, determining the model's capacity.
- **Dropout Probability:** Provides regularization to prevent overfitting.
- **Output Dimension:** Set to 3, to predict the odds for the Home, Away, and Draw outcomes.

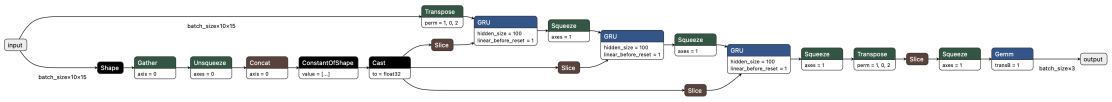

Figure 4.6: GRU Model Architecture

Both architectures, LSTM and GRU, provide a deep learning framework that can capture intricate temporal dependencies in football match data. Their ability to leverage sequential patterns makes them particularly suitable for predicting odds, helping bettors gain an edge in the competitive sports betting market.

4.5. MODEL TRAINING AND VALIDATION

Training a model is a crucial step in any machine learning workflow. For our football betting application, this process becomes all the more critical, as the accuracy and robustness of our predictions can have direct financial implications. The following section walks through the process of training our Stacked LSTM network.

4.5.1. Hyperparameters and Configuration

The training process is guided by a series of hyperparameters, which are pre-defined settings that dictate how the model learns. The following hyperparameters were chosen for our training:

- Number of Epochs: An epoch represents a full cycle through the training data.
- Loss Function: Mean Squared Error Loss (MSE) - This function measures the average squared difference between the estimated odds and the actual odds.
- Optimizer: AdamW - An extension to the Adam optimizer with weight decay, which helps in regularizing the model.
- Learning Rate: Initially set to 0.001 with a maximum of 0.005, controlling the step size during optimization.

To adaptively adjust the learning rate during training, a *OneCycleLR* learning rate scheduler was employed. This scheduler dynamically alters the learning

rate, allowing the model to converge faster and potentially attain better performance.

4.5.2. Training Procedure

The model training process begins by sending the LSTM model to the appropriate device (typically a GPU for faster computation). For each epoch, the model is set to training mode, and data is fed in batches. After predicting the outcomes using the LSTM, the predictions are compared to the true values using the MSE loss. The optimizer then updates the model's weights based on the computed gradients.

During training, the learning rate is adaptively adjusted using the scheduler. After processing all batches for an epoch, the model's performance is evaluated on a validation set to gauge its generalization capability. This process is repeated for all epochs, with the model continually refining its weights.

4.5.3. Data Splitting

The models, being a type of recurrent neural network (RNN), inherently work with sequences. In the context of our football betting application, the sequences represent time windows of betting data. Given the irregular time intervals at which data packets arrive from the raw historical files and the listener, it's crucial to standardize this for model training.

To this end, we construct sequences using a fixed window of 10 seconds for each runner. Each such sequence serves as an input to our model, which then attempts to predict the odds for the next 10-second interval. This sequential arrangement captures the temporal dependencies in the odds movement and provides our model with the necessary context to make accurate predictions.

For each prediction task, the model is supplied with 10 intervals, each spanning 10 seconds, aggregating to a total of 100 seconds of historical data. This 100-second window provides the model with sufficient context to capture the temporal dynamics and nuances in the odds movement. Using this substantial context, the model is then tasked with forecasting the odds for the subsequent 10-second interval. By structuring the data in this manner, we ensure that the model is well-equipped with a rich history, enhancing its capability to make informed and accurate predictions for the immediate future.

To better understand the data distribution for training, validation, and testing, consider the following breakdown for a single market:

Dataset	Percentage of Total Data	Purpose
Training	75%	Model training
Validation	15%	Hyperparameter tuning
Testing	10%	Model evaluation

Table 4.1: Split of a single historical/live market data for LSTM model

The rationale behind this split is to ensure that a significant portion of the data is

used for training, enabling the model to learn the intricate patterns in the data. The validation set aids in tuning and ensuring that our model isn't overfitting to the training data. Finally, the test set, being the most recent 10% of the data, allows us to assess the model's performance in a realistic, forward-looking scenario, mimicking how the model would perform in real-time betting situations.

Data Pre-processing and Feature Generation

While a detailed discussion on data pre-processing and feature generation is provided in the previous section, it's worth noting here that our LSTM model was trained on both raw and synthetic features. Using functions like ***calculate_roc***, ***calculate_moving_average***, and ***calculate_macd***, we transformed our raw data into informative features that capture the dynamics of football odds and games.

4.5.4. Training Strategy

Model training is an art as much as it is a science. The strategy adopted for training can significantly influence the model's performance:

- **Grid Search for Hyperparameter Tuning:** We utilized grid search to systematically find the optimal hyperparameters for our models, enhancing their performance. This approach automates the process of selecting the best parameters that govern the learning process.
- **One-Cycle Learning Rate:** The learning rate dictates the steps the model takes to adjust its weights during training. Instead of using a constant learning rate, we adopted a one-cycle policy that starts with a small learning rate that gradually increases for the first half of training and then decreases for the second half. This dynamic adjustment aids in faster convergence and prevents the model from getting stuck in local minima.
- **Early Stopping:** Early stopping was implemented to monitor the model's validation performance and halt training if it started to degrade. This strategy helps in preventing overfitting, ensuring that the model generalizes well to new, unseen data.

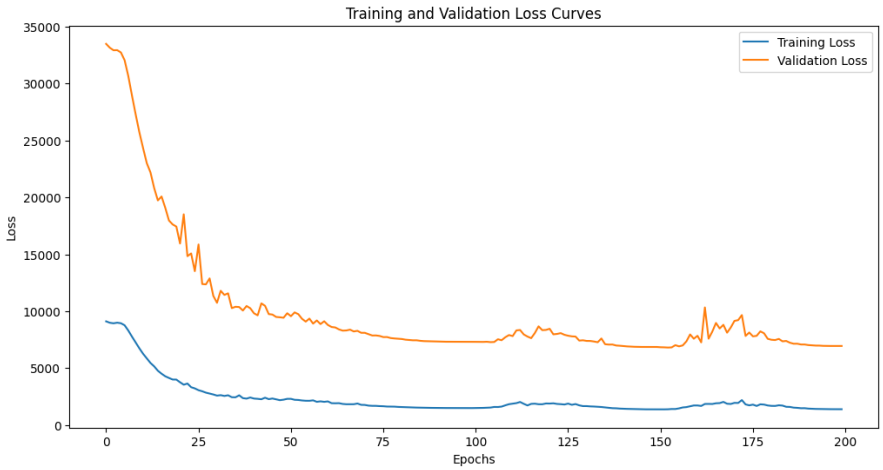

Figure 4.7: Example of GRU model Training

4.6. MODEL EVALUATION AND RESULTS

In the intricate domain of sports analytics and betting, evaluating the performance of deep learning models is crucial for predicting in-play betting odds in football. For this experiment, we used preprocessed Betfair exchange data enriched with momentum-based features such as Rate of Change (RoC), Moving Average (MA), Moving Average Convergence Divergence (MACD), and Signal line. The raw data was first transformed into a fixed window size of 10 seconds and then scaled using Min-Max scaling to ensure uniformity and to facilitate model training, as discussed earlier.

For the training process, we employed **Grid Search** to fine-tune the hyperparameters for the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. To optimize the learning rate during the training, the **One-Cycle Learning Rate policy** was used. Additionally, **Early Stopping** was integrated into the training process to prevent the model from overfitting and to ensure the most optimal model is selected for evaluation.

The results of the trained models are particularly insightful when it comes to predicting the three possible outcomes of a football match: Home win, Away win, and Draw. These outcomes are represented as odds in the betting market, and our models aim to predict these odds as closely as possible to the real market odds. Inspired by financial markets and algorithmic trading, these deep learning models offer a nuanced approach to sports betting. By predicting these three key odds, the models provide a complete spectrum of betting opportunities for a given match. The predicted odds serve not just as an indicator of which team is likely to win or if a draw is probable, but also offer quantifiable measures of the confidence level associated with each prediction. This comprehensive approach enables a nuanced understanding of match dynamics, thereby providing bettors with actionable insights for formulating more effective betting strategies.

After training both the LSTM and GRU models, we assessed their capabilities using a range of evaluation metrics. These metrics include R-squared, which measures the proportion of the variance for the dependent variable that's explained by the independent variables; Explained Variance Score, which gauges the model's ability to capture the underlying data distribution; and Mean Squared Error (MSE), which quantifies the average squared differences between the observed actual outcome values and the values predicted by the model.

By using these rigorous evaluation metrics, we offer a comprehensive view of each model's performance, thereby allowing us to make informed decisions for real-world football betting strategies.

4.6.1. Results of LSTM Model

The Long Short-Term Memory (LSTM) model displayed remarkable performance in predicting the odds for the three possible outcomes in football matches: Home win, Away win, and Draw. The visual analysis, which will be presented below, corroborates the model's capabilities. It shows how the LSTM model's predictions

for each of the test data selections align closely with the real market odds over time.

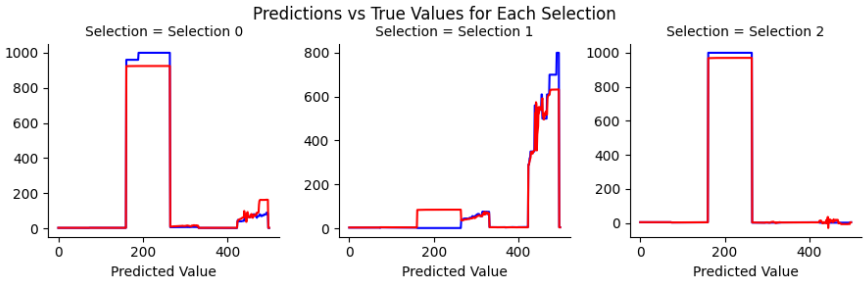

Figure 4.8: Temporal predictions of the LSTM models for the test data selections. It provides insights into how closely the predictions match the actual outcomes over time.

Furthermore, the residuals plot, which illustrates the differences between the model's predictions and the actual odds, further supports the effectiveness of the LSTM model. The residuals are consistently close to zero, indicating that the model's predictions are accurate.

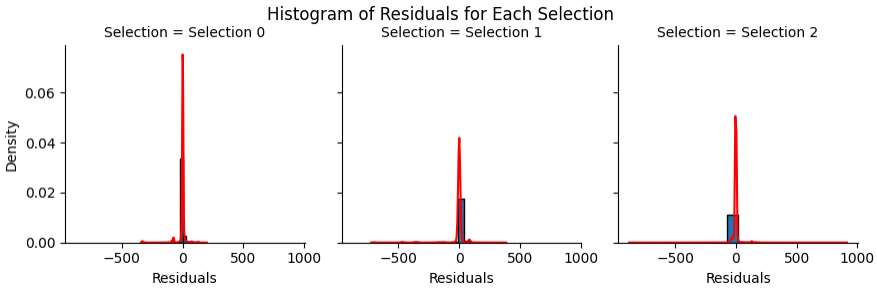

Figure 4.9: Residual plots showcase the differences between the actual and predicted values for the LSTM models. Patterns in residuals can hint at potential improvements in the modeling process.

The model's performance can also be quantified using several metrics:

Metric	Value
R-squared	0.9322
Explained Variance Score	0.9324
Mean Squared Error	46.55

Table 4.2: LSTM Model Metrics

The high R-squared value of 0.9322 and an Explained Variance Score of 0.9324 suggest that the LSTM model is highly accurate. The low Mean Squared Error of 46.55 further confirms this.

4.6.2. Results of GRU Model

The Gated Recurrent Unit (GRU) model, while not as effective as the LSTM model, still offers reasonably good performance in its predictions. The visual

analysis is indicative of this, displaying the model's predictions and how they compare to the actual odds over a period.

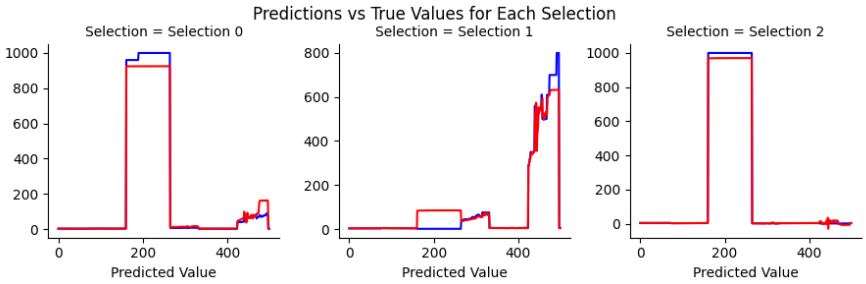

Figure 4.10: Temporal predictions of the GRU models for the test data selections. It provides insights into how closely the predictions match the actual outcomes over time.

The residuals plot for the GRU model further highlights the differences between the model's predictions and the actual odds, helping us identify areas where the model might need further tuning.

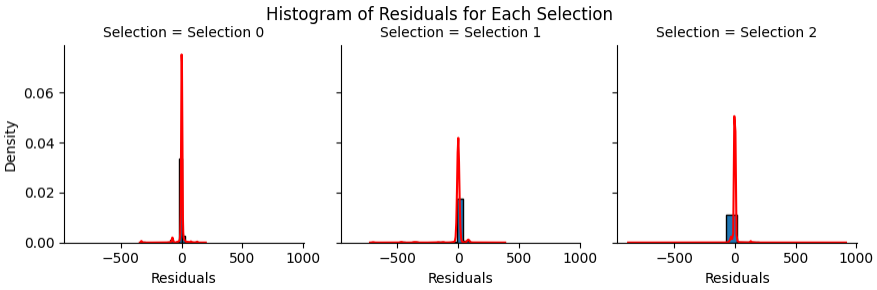

Figure 4.11: Residual plots showcase the differences between the actual and predicted values for the GRU models. Patterns in residuals can hint at potential improvements in the modeling process.

The model's quantitative metrics are as follows:

Metric	Value
R-squared	0.7856
Explained Variance Score	0.7900
Mean Squared Error	82.81

Table 4.3: GRU Model Metrics

The R-squared value of 0.7856 and Explained Variance Score of 0.7900 suggest that the GRU model is relatively accurate, but not as much as the LSTM model. The higher Mean Squared Error of 82.81 also supports this conclusion. Both the visual analysis and these metrics are crucial for identifying the most suitable model for our application, and in this case, the LSTM model appears to be the more accurate choice.

4.7. ANALYSIS AND DISCUSSION

The results presented in the preceding sections provide a basis for a detailed discussion about the predictive capabilities of our deep learning models, their applicability in in-play betting strategies, and the unique challenges presented by in-play data. In this section, we delve into the analytics to draw insights and discuss implications.

4.7.1. High Variability in Predictive Performance

The nature of in-play betting presents a high degree of variability and noise, especially given the 10-second window for last traded prices. This has a consequential impact on our LSTM and GRU models, making the prediction landscape considerably more challenging than pre-game models. Despite the implementation of a simulation framework to test betting strategies, the extremely high noise levels in the data make the development and implementation of betting strategies ineffective, yielding very poor results.

4.7.2. Infrastructure and Reliability Considerations

In-play betting demands a robust and highly available infrastructure. The system must be designed to handle multiple betting markets concurrently and to recover quickly and gracefully in case of failures, limiting exposure and closing any open trades. The infrastructure costs for maintaining such a high-availability system are non-trivial and must be factored into the overall profitability calculations for the betting strategies employed.

4.7.3. Drift in Match Outcomes Over Time

In football, especially in the context of in-play betting, understanding the dynamics of odds drift can provide an edge. Drift refers to the gradual movement of odds in a particular direction over time. For instance, if a match between Team A and Team B is 0-0 and stays that way as the clock ticks, the odds for a draw will typically shorten (become less profitable), while the odds for either team winning will generally drift (become more profitable).

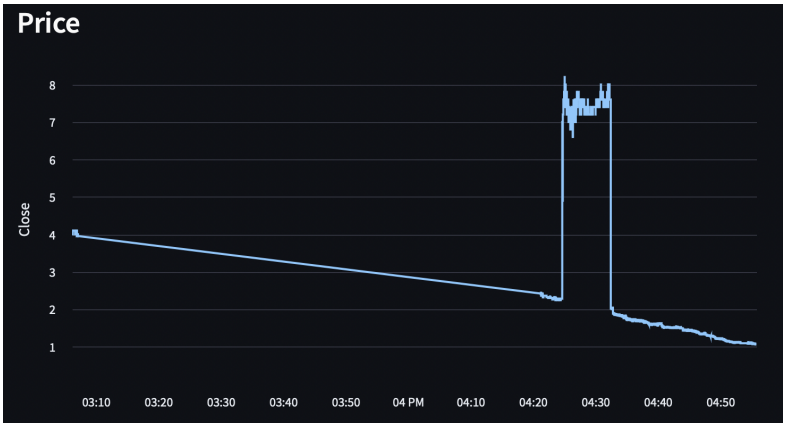

Figure 4.12: Drift of draw odds over time during the Arsenal vs. Brentford match on 11-02-2023, reflecting the game's progression to a 1-1 scoreline.

This drift is not random but rather a reflection of the reducing time available for a new event (like a goal) to change the outcome of the match. In a way, it's a race against time, and the odds are adjusted accordingly to reflect this. Understanding this drift can provide punters with opportunities for "timing" their bets to maximize potential returns. For example, if the odds on a draw are shortening, a timely bet could lock in higher potential profits than if the same bet were made later in the game.

The drift is particularly noticeable after key events such as goals, red cards, or injuries, where the market re-adjusts its expectations for all three outcomes (Home win, Away win, Draw). This re-adjustment period can present profitable opportunities for traders who understand the dynamics of odds drift and can predict how the market will settle post-event.

In the context of our models, these patterns seem to be captured and effectively leveraged, this could offer another layer of sophistication to the betting strategies. Given the fixed time frame of a football match, these drift patterns present an additional edge in the highly competitive landscape of in-play betting that our models are able to capture.

4.7.4. Adaptability and Risk Management

Given the dynamic nature of in-play betting, strategies that can adapt in real-time to changing game conditions and odds offer a potential advantage. Our results reflect the need for such adaptability, particularly in the face of high data variability and infrastructure demands. Risk management, which includes quick failure recovery and limited exposure, becomes crucial in this volatile environment.

4.7.5. Conclusion

Our exploration into the realm of in-play betting and predictive modeling for football match outcomes has revealed the complexity and dynamic nature of the field. While the use of deep learning models like LSTMs and GRUs has shown promise, the challenges associated with the high variability of in-play data and the need for a robust, high-availability infrastructure cannot be overlooked. This demonstrates that such trading strategies require expensive infrastructure to execute. Further, on failure systems must be able to limit the exposure and close any open positions preventing drastic losses.

The experiment underscores the necessity of a multi-faceted approach. From meticulously tailored strategies to rigorous validation methods and from infrastructure design to risk management, each element plays a crucial role in the overall effectiveness and profitability of in-play betting strategies.

However, it's important to remain aware of the inherent unpredictability that comes with sports and betting. While machine/deep learning offers sophisticated tools for analysis and prediction, it's not a silver bullet for guaranteed profits. The path forward should be one of responsible betting, continuous learning, and agile adaptation to the ever-changing dynamics of football matches and betting markets.

4.8. SUMMARY

In this research chapter, we ventured into the dynamic and uncertain world of in-play football betting, tackling its complexities with deep learning models. Notably, we developed a listener to collect live in-game data from Betfair, an essential element for our real-time predictive models. This listener is equipped with features like scheduling and the capability to handle multiple markets and games simultaneously, making it a cornerstone in our research framework.

Inspired by financial markets, we employed additonal synthetic features like Rate of Change, Moving Average, and Moving Average Convergence Divergence for short-interval predictions. Our approach was deeply rooted in research based methodologies and practices, making it relevant and rigorous.

One of the defining aspects of our work was the creation of a real-time simulation framework. This tool serves not only as a mechanism for model evaluation but also as a testing ground for various in-play betting strategies. It is designed to operate under high-availability conditions and could be configured to be resilient to failures, reflecting the practical constraints of in-play betting.

In summary, this chapter bridges machine learning, real-time analytics, and sports betting into an integrated research endeavor. Beyond the immediate insights for in-play football betting, our lasting contribution is the robust, extendable framework, including the live data listener, that sets a solid foundation for future research and practical applications in this compelling domain.

Chapter 5

Conclusion

This final chapter summarizes the research journey that has been undertaken providing an overview of the main findings and methodologies used throughout the thesis. It highlights the contributions made to the areas of sports analytics and machine learning emphasizing how these fields intersect, in the context of football betting. The wider impact of this research its potential to transform the sports betting industry is discussed, emphasizing how the practical insights gained can be applied in real world settings. Additionally this chapter discusses directions for exploration for further research, in this ever evolving field.

5.1. SUMMARY

The two experiments in our study offer a comprehensive comparison of predictive modeling in football betting, focusing on pre-game and in-play strategies. The key takeaways from this juxtaposition are illuminating.

Firstly, the ***complexity and overhead*** requirements differ substantially between the two experiments. Experiment 1 shows that straightforward machine learning models, coupled with simple betting strategies, can yield promising results with minimal infrastructure requirements. This reduces both the complexity and the overhead costs, making it a more accessible option for most punters. On the other hand, Experiment 2 demands a robust, high-availability infrastructure to cater to the real-time aspects of in-play betting, increasing both the complexity and the cost.

Regarding ***predictive accuracy***, Experiment 1 employs models like Logistic Regression, Gradient Boosting, and Random Forest, which demonstrated high levels of accuracy, especially for the English Premier League. Conversely, the deep learning models used in Experiment 2 struggle with the high variability and noise in in-play data, making the prediction landscape considerably more challenging and the results less reliable.

In terms of strategic betting, Experiment 1 enabled profitable strategies to be developed, such as the Dynamic Fractional Kelly Criterion, especially for markets like the English Premier League. Experiment 2, however, exposed the difficulties

in translating these strategies to the in-play betting scene due to the inherent volatility in real-time data.

Profitability also seems to favor Experiment 1. The simpler models and strategies appear to yield more predictable outcomes, potentially leading to profits. The added costs and complexities of the infrastructure needed for in-play betting in Experiment 2 could potentially offset any gains, making it a riskier venture.

Lastly, ***risk and variability*** are inherently higher in Experiment 2 due to the volatile nature of in-play data and the infrastructure demands. Experiment 1, with its simpler models and strategies, offers more predictable and less risky outcomes.

In summary, while both experiments demonstrate the challenges and potentials of predictive modeling in football betting, Experiment 1 emerges as the more accessible, less risky, and potentially more profitable avenue. Experiment 2, although intriguing, serves as a cautionary tale about the complexities and risks associated with in-play betting. Therefore, for those seeking a less complex and lower-risk venture with potential profitability, the pre-game models and strategies of Experiment 1 appear to be the more prudent choice.

5.2. CONTRIBUTIONS

The culmination of this research journey has not only expanded the horizons of sports analytics and machine learning but has also provided practical tools and insights that can be harnessed in the real-world context of sports betting. This thesis has yielded several pivotal contributions, each of which is outlined below:

1. **An Extensive Framework for Predictive Modelling in Football:** At the heart of this thesis lies the development of a meticulous framework tailored for predictive modelling in football. This framework, grounded in rigorous machine learning methodologies, serves as a blueprint for constructing, evaluating, and iterating on prediction models. Its systematic nature ensures scalability and adaptability, positioning it as an invaluable asset for future academic explorations and real-world applications in the realm of football predictions.
2. **Exploration and Comparison of Various Betting Strategies:** Venturing beyond mere predictions, this research delved deep into the world of betting strategies. Through extensive experimentation and analysis, a spectrum of strategies, from simple to sophisticated, was explored. The comparative analysis brought forth nuanced insights, revealing the interplay between the strategies and their corresponding predictive models. This exploration provides a roadmap for selecting and implementing betting strategies that align with specific predictive models, optimizing potential returns.
3. **In-Play Momentum Betting using Stacked LSTM Networks:** Marking a significant departure from traditional techniques, this research introduced the innovative application of stacked Long Short-Term Memory (LSTM)

networks for in-play momentum betting. By harnessing the power of deep learning, the research bridged the gap between advanced machine learning and sports betting, paving the way for future innovations in this interdisciplinary domain.

4. **Robust Backtesting Framework and Data Listener for In-Play Football Data:** Recognizing the importance of real-time data in the dynamic world of in-play betting, the thesis introduced a robust backtesting framework complemented by a state-of-the-art data listener. This dual toolset not only facilitates the seamless collection and analysis of in-play football data but also ensures that betting strategies are tested and refined in a realistic environment. This contribution stands as a testament to the practical orientation of the research, underscoring its value for both researchers and industry practitioners.
5. **Comparative Analysis of Value Betting and Momentum Betting Strategies:** In a bid to provide a holistic view of the betting landscape, the research undertook a comprehensive comparative analysis between value betting and momentum betting strategies. This analysis demystified the strengths, limitations, and applications of each strategy, shedding light on their suitability across diverse match scenarios and market dynamics. The insights derived from this analysis offer a nuanced understanding, guiding punters in making informed betting decisions.

The contributions outlined above not only enrich the academic discourse in sports analytics and machine learning but also hold significant practical implications, especially for stakeholders in the sports betting industry.

5.3. FUTURE WORK

While this thesis has laid the foundation for predictive modelling in football and the subsequent application of betting strategies, there remain myriad avenues for further exploration and refinement. This section outlines potential enhancements and new directions for future research:

1. **Value Betting Experiment Enhancements:**
    - *Incorporating Richer Pregame Data:* The predictive prowess of the model could be significantly enhanced by incorporating a richer set of pregame data. Specifically, integrating player performance metrics, recent form, injuries, and other granular data points can provide a more comprehensive view of the match context, thereby improving prediction accuracy.
    - *Deriving Advanced Synthetic Features:* While the current set of synthetic features has proven valuable, there is potential for deriving even more informative features. Delving into advanced statistical methodologies or leveraging domain-specific insights could yield features that capture nuanced interactions and patterns, enhancing the model's ability to predict outcomes.

- *Incorporating Game Tactics:* Understanding team formations and tactical approaches can provide insights into the likely outcome of a match. Integrating data about team tactics and strategies could refine predictions.
- *Weather and Crowd Influence:* External factors, such as weather conditions and crowd presence (home advantage), can significantly impact match outcomes. Incorporating such data might offer additional predictive power.

2. **Momentum Betting Experiment Enhancements:**
    - *Exploring Alternative Models:* While stacked LSTMs have showcased their capability in modeling momentum betting, other advanced models could be explored. Potential candidates include Convolutional Neural Networks (CNNs) for detecting patterns in time series data, Transformer-based models like BERT for sequence prediction, and Attention Mechanisms to weigh the importance of different time steps.
    - *Integrating Pregame State Data for In-Play Predictions:* A promising direction for future research is the fusion of pregame data (from Experiment 1) with in-play data (from Experiment 2). By creating a holistic model that leverages both types of data, one can potentially achieve superior predictive accuracy, especially in the early stages of a match when in-play data is sparse.
    - *Event-Driven Predictions:* Building models that specifically predict the likelihood of certain in-game events, such as goals, penalties, or red cards, using event-driven architectures.
    - *Anomaly Detection for Betting Opportunities:* Deploying anomaly detection techniques to identify rare but highly profitable betting opportunities in real-time, especially when the odds do not reflect the in-play dynamics.

In addition to the aforementioned directions, a broader exploration of betting markets, beyond football, could provide insights into the generalizability of the presented methodologies. Furthermore, integrating real-time feedback mechanisms, where the model learns from its own predictions and the actual outcomes, could lead to adaptive models that evolve with the ever-changing dynamics of sports matches.

The horizon of sports analytics and machine learning-based betting strategies is vast, and this thesis merely scratches the surface. It is hoped that the foundations laid herein will inspire and guide future researchers in pushing the boundaries even further.

References

Aldridge, Irene (2013). *High-frequency trading: a practical guide to algorithmic strategies and trading systems*. John Wiley & Sons.

Brown, A. and F. Yang (2004). “The Influence of Contextual Factors and Decision Familiarity on Consumer Decision Making in Physical and Virtual Stores”. In: *Journal of Interactive Marketing* 18.4, pp. 30–43.

Bunker, R. and F. Thabtah (2019). “A machine learning framework for sport result prediction”. In: *Applied Computing and Informatics*.

Cho, Kyunghyun et al. (2014). “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078*.

Chong, E., C. Han, and F. C. Park (2017). “Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies”. In: *Expert Systems with Applications* 83, pp. 187–205.

Clarke, S. R. and J. M. Norman (1995). “Home ground advantage of individual clubs in English soccer”. In: *The Statistician*, pp. 153–161.

Constantinou, A. C. and N. E. Fenton (2012). “Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models”. In: *Journal of Quantitative Analysis in Sports* 8.1.

— (2013). “Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries”. In: *Journal of Quantitative Analysis in Sports* 9.1, pp. 37–50.

Divos, Peter et al. (Nov. 2018). “Risk-Neutral Pricing and Hedging of In-Play Football Bets”. In: *Applied Mathematical Finance* 25, pp. 1–21. DOI: 10.1080/1350486X.2018.1535275.

Dixon, M. and S. Coles (1997). “Modelling Association Football Scores and Inefficiencies in the Football Betting Market”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46.2, pp. 265–280.

Dixon, M., D. Klabjan, and D. Bang (2017). “Classification-Based Financial Markets Prediction Using Deep Neural Networks”. In: *Algorithmic Finance* 6.3-4, pp. 67–77.

Elo, Arpad E. (1978). *The Rating of Chess Players, Past and Present*. Arco.

Fátima Rodriguesa, Ângelo Pinto (2022). “Prediction of football match results with Machine Learning”. In: *International Conference on Industry Sciences and Computer Science Innovation* 204, pp. 463–470.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.

Hubáček, Ondřej, Gustav Šourek, and Filip Železný (2019). “Learning to predict soccer results from relational data with gradient boosted trees”. In: *Machine Learning* 108.1, pp. 29–44.

Humphreys, B. R. and L. Perez (2012). “Who Bets on Sports? Characteristics of Sports Bettors and the Consequences of Expanding Sports Betting Opportunities”. In: *Journal of Economic Behavior & Organization* 30.2, pp. 579–598.

Hvattum, L. M. and H. Arntzen (2010). “Using Elo ratings for match result prediction in association football”. In: *International Journal of Forecasting* 26.3, pp. 460–470.

Karlis, D. and I. Ntzoufras (2003). “Analysis of sports data by using bivariate Poisson models”. In: *The Statistician* 52.3, pp. 381–393.

Kelly, J. L. (1956). “A New Interpretation of Information Rate”. In: *Bell System Technical Journal* 35, pp. 917–926.

MacLean, L. C., W. T. Ziemba, and G. Blazenko (1992). “Growth Versus Security in Dynamic Investment Analysis”. In: *Management Science* 38.11, pp. 1562–1585.

Pollard, Richard and G. Pollard (2005). “Home advantage in soccer: A review of its existence and causes”. In: *International Journal of Soccer and Science Journal* 3.1, pp. 28–38.

Rue, H. and O. Salvesen (2000). “Prediction and retrospective analysis of soccer matches in a league”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 49.3, pp. 399–418.

Smith, M. A., D. Paton, and L. Vaughan Williams (2009). “Do bookmakers possess superior skills to bettors in predicting outcomes?” In: *Journal of Economic Behavior & Organization* 71.2, pp. 539–549.

Szymanski, S. and T. Kuypers (2003). *Winners and Losers: The Business Strategy of Football*. Penguin UK.

Thorp, E. (1966). “The Kelly Criterion in Blackjack Sports Betting, and the Stock Market”. In: *Handbook of Asset and Liability Management* 1, pp. 385–428.

Vaughan Williams, L. (2014). *Information efficiency in financial and betting markets*. Cambridge University Press.

Vlastakis, N., G. Dotsis, and R. N. Markellos (2009). “How Efficient is the European Football Betting Market? Evidence from Arbitrage and Trading Strategies”. In: *Journal of Forecasting* 28.5, pp. 426–444.

Zhang, Qiyun et al. (2022). “Sports match prediction model for training and exercise using attention-based LSTM network”. In: *Digital Communications and Networks* 8.4, pp. 508–515. ISSN: 2352-8648. DOI: https://doi.org/10.1016/j.dcan.2021.08.008.