POLITECNICO DI TORINO

Master's Degree in Computer Engineering

Master's Degree Thesis

Enhancing Prediction Accuracy in Low-Scoring Games through Computational Intelligence and Mathematical Models

Supervisors

Giovanni SQUILLERO

Alberto TONDA

Candidate

Jacopo TARAMASSO

OCTOBER 2024

Summary

The aim of this thesis is to design an intelligent predictive model for low-scoring competitive games. More specifically, this model integrates classical mathematical approaches with advanced computational intelligence techniques to leverage historical datasets for more accurate predictions.

The thesis is divided into four main parts:

- The first section provides an in-depth introduction to the theoretical background, covering essential concepts in game theory and reviewing the most prominent models in the existing literature. This part lays the foundation to better understanding the challenges and methodologies used in forecasting low-scoring outcomes.

- The second part focuses on the creation and the refinement of the dataset. It details the process of data collection, the criteria for selecting relevant features and the methodologies used to ensure the dataset is robust and representative.

- The third section is dedicated to the development of the predictive model. It covers the introduction of new features, the selection of the most appropriate model and the optimization of its parameters; some of which are taken from the literature of existing models, while others are optimized by original research and experimentation.

- The final section applies the model to various datasets, evaluating its performance using a range of metrics. The results are then analyzed, comparing the model's predictions to actual outcomes. The thesis concludes with a discussion on the model's effectiveness, potential limitations, and suggestions for future improvements and extensions.

Data beats emotions

Table of Contents

List of Tables

List of Figures

Glossary and Acronyms

Efficient Market Hypothesis (EMH)

The Efficient Market Hypothesis is a financial theory that suggests that asset prices fully reflect all available information. According to this hypothesis, it is impossible to consistently achieve higher returns than the overall market through stock picking or market timing because market prices always incorporate and respond to information instantly.

Prediction Markets (PMs)

Prediction markets are speculative markets where participants trade contracts based on the outcomes of future events. These markets leverage the collective knowledge and predictions of participants, often resulting in highly accurate forecasts. Participants can trade contracts tied to political events, economic trends, and more, with prices reflecting the probability of specific outcomes.

Yield

Yield is a metric used in economics and other fields to indicate the return on one or more investments over a specific period of time. It is typically calculated by dividing the profit or income generated by the investment by the total value of the investment, resulting in a percentage that represents the return. A positive yield indicates a profit, while a negative yield indicates a loss.

Vig

In sports betting, the vig is the fee that sportsbooks charge, embedded in the odds to ensure a profit regardless of the outcome. The implied probability from these odds exceeds 100%, accounting for the vig, which means the probability reflected by the bookmaker is less favorable than the true likelihood of the event. This difference ensures the sportsbook's edge.

GDP

GDP is the total monetary value of all goods and services produced within a country's borders over a specific period, typically measured annually or quarterly. It is a key indicator used to assess the economic health of a country, reflecting the size and strength of its economy. GDP is often categorized into nominal GDP (measured at current market prices) and real GDP (adjusted for inflation). It is commonly used to compare the economic performance of different countries or to track economic growth over time.

Chapter 1

Introduction

1.1 Forecasting and Game Theory

Forecasting plays a critical role in decision-making across various domains, from business and economics to sports and technology. The ability to predict future events [1] based on historical data and patterns empowers individuals and organizations to anticipate changes, optimize strategies, and mitigate risks. This is especially relevant in environments where outcomes depend on the behavior of multiple agents or variables, such as it happens in Game Theory [2].

In the context of this thesis, Game Theory provides a valuable framework for understanding how individuals or systems can anticipate and react to the choices made by others, leading to create a dynamic interplay of choices. Similarly, in forecasting, a model must take into consideration the influence of external factors and potential fluctuations in the data to achieve accurate predictions, creating an additional layer of depth. Both disciplines share a common goal: to maximize outcomes by analyzing patterns, understanding behaviors, and preparing for a range of possible scenarios. Just as Game Theory helps participants anticipate and respond to others' strategies, forecasting gives decision-makers the ability to anticipate future conditions and adjust their strategies accordingly. In this way, forecasting becomes a powerful tool for making well-informed choices, minimizing risks, and taking advantage of emerging opportunities in an increasingly uncertain world. Through this connection between forecasting and strategic decision-making, the power of prediction extends beyond the numbers, shaping practical outcomes in real-world scenarios.

In this competitive landscape, the game itself becomes one among predictive models, where the model that can best anticipate future outcomes wins. It is no longer just about understanding the past but overcoming other models in accuracy and adaptability. Success belongs to the model that can predict the outcomes the most accurately, much like in Game Theory, where strategic foresight determines the winner.

1.2 Goal of the model

The primary goal of this thesis is to develop a robust prediction model specifically designed to forecast low-scoring games, with a particular focus on soccer. While soccer serves as the primary example in this study, the model's framework can be adapted to other low-scoring sports as well. The aim is to achieve greater accuracy than the predictions offered by major bookmakers, who often struggle to account for all the nuances of such games, expecially in minor leagues.

In soccer, accurately predicting outcomes is especially challenging due to the sport's low-scoring nature, where a single event can drastically alter the result. The variability in team performance, tactical decisions, and external conditions, such as weather or injuries, adds complexity to the prediction process.
The model here developed seeks to harness these variables to pinpoint key indicators that influence outcomes, exploiting potential inefficiencies in bookmaker odds.

The main architecture of the predictor is designed to be data-driven, leveraging historical data to identify patterns, trends, and correlations that can achieve more accurate predictions. By integrating classical mathematical approaches with advanced optimization algorithms, the model aims to provide a sharper edge in forecasting low-scoring games, offering valuable insights into predictive modelling in sports and beyond.

1.3 State-of-the-Art on Modelling Association Football Scores

The modelling of football scores has long been a topic of debate among statisticians, primarily due to the complexity and perceived randomness of goal scoring. Early models, such as the Poisson distribution [3], provided a simple framework for predicting number of goals but struggled to capture the nuances of team-specific strengths and weaknesses. The Poisson model is well-suited for predicting events that happen independently and with a constant probability over time-conditions that can be scarcely applied to football scoring, where goals are relatively rare and spread throughout the match.

The key idea is that each team's number of goals can be treated as a random variable following a Poisson distribution, with the expected number of goals (λ) being based on the team's offensive strength and the opponent's defensive capabilities.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where:

- $P(X = k)$ is the probability of observing k goals
- λ is the expected number of goals

For example, if the expected goals (λ) for a match between two teams are known, the Poisson distribution can be used to calculate the probabilities of different scorelines. This is commonly applied in betting markets to estimate the odds of outcomes such as the total number of goals, win/draw/lose probabilities, or exact scorelines.
However, the raw Poisson model has significant limitations. It assumes that the mean and variance of goal distributions are equal, which often isn't the case in real matches where variances tend to be larger, leading to deviations from the Poisson fit. Additionally, the model doesn't take into consideration team-specific factors such as home advantage or varying levels of team quality, which are crucial in football. This led statisticians to seek for more advanced models to better capture the complexity of football scoring dynamics.

Early works by Moroney (1951) and Reep et al. (1971) [4] rejected the Poisson model in favor of the Negative Binomial distribution. These early models argued that football matches are dominated by chance events, making it difficult to fit goal distributions using a simple Poisson framework. Moroney showed that the Negative Binomial distribution [5] provided a better fit, particularly because the variance in football scores was higher than the Poisson model would predict.
Reep, Pollard, and Benjamin extended this analysis to other ball games, suggesting that chance plays a dominant role in these sports, often overriding the inherent skill differences between teams. However, this view was challenged by Hill (1974) [6], who demonstrated that football experts could successfully predict league outcomes at the beginning of a season, implying that skills, rather than chance, play a larger role in a long time scenario. This led to the awareness that while chance events can influence individual matches, skill differences among teams become evident over multiple games.

As a result of these observations, Maher (1982)[7] revisited the Poisson model with the intention of exploring whether it could still be useful in describing football scores, given the inherent qualities of teams such as attack and defense.
Maher proposed that previous rejections of the Poisson model might have been premature and sought to incorporate team-specific characteristics into the model to better capture the nuances of football scoring.

1.3.1 The Maher Model

Maher's model introduces a more sophisticated approach to Poisson modelling by accounting for the attacking and defensive strengths of teams, both at home and away. The idea is that football scores are not purely at random but are influenced by team characteristics such as offensive prowess and defensive weaknesses.
Maher suggests that the number of goals scored by each team can be treated as independent Poisson variables, with the mean of these variables reflecting the team's strength in attack and the opposing team's weakness in defense.

In the case where team i plays at home against team j, Maher defines:

- X_{ij} as the number of goals scored by team i *(home team)*
- Y_{ij} as the number of goals scored by team j *(away team)*

These variables follow independent Poisson distributions with means:

$$\lambda_{ij} = \alpha_i \beta_j$$

Where:

- α_i represents the attacking strength of team i at home
- β_j represents the defensive weakness of team j when playing away

$$\mu_{ij} = \gamma_i \delta_j$$

Where:

- γ_i represents the defensive weakness of team i at home
- δ_j represents the attacking strength of team j when playing away

Maher estimates these parameters using maximum likelihood estimation, fitting the model to historical data to determine the best values for α, β, γ, and δ.
Through likelihood ratio tests, Maher shows that models allowing for team-specific parameters significantly improve the fit compared to the simpler, restricted models. Specifically, models that allow teams to have different attack strengths and defenses (both home and away) fit the data best.
However, Maher notes that while the Poisson model fits well overall, it assumes that team scores are independent, which may not always be true. In a football match, teams adjust their strategies based on the current score. For instance, a team that is losing may take more risks defensively to score, which increases the likelihood of conceding additional goals.
Considering this, Maher introduces a bivariate Poisson model, which allows a correlation between the scores of the two teams. The correlation coefficient is estimated to be approximately 0.2, indicating that the scores of home and away teams are not entirely independent. The bivariate model improves the fit, especially when predicting the frequency of drawn matches (where both teams score the same number of goals).

Maher's work shows that, contrary to earlier studies, the Poisson model can be an effective tool for modelling football scores when team-specific strengths and weaknesses are considered. While the independent Poisson model provides a reasonable fit, the inclusion of a correlation factor through the bivariate Poisson model yields even better results. This suggests that while chance plays a role in a small number of matches, team quality and strategy are important factors over the course of a season.

The study is significant because it offers a relatively simple yet effective framework for predicting football outcomes, moving away from the over-complication of the Negative Binomial model. Maher's approach, with its focus on team-specific parameters, has influenced subsequent research in sports analytics and remains a foundational model in the statistical analysis of football scores.
Following Maher's pioneering work in 1982, various other models have emerged in an attempt to improve predictive accuracy in football outcomes. Among these, the model introduced by Dixon and Coles (1997) [8] is one of the few that can be seen as a direct continuation and enhancement of Maher's framework.

1.3.2 Dixon-Coles Model

Dixon and Coles recognized the potential of Maher's[7] independent Poisson model but aimed to refine it further, particularly for its application in betting strategies. Their model builds directly on Maher's assumption that the number of goals scored

by both the home and away teams can be modeled as independent Poisson variables. However, Dixon and Coles introduced key modifications to address the limitations of Maher's model, especially in capturing the fluctuating performance of teams over time.
The Dixon-Coles model begins with the same fundamental assumption as Maher's model: the number of goals scored by the home team X_i and the away team Y_j are independent Poisson-distributed random variables. This assumption allows for each team to have separate attack and defense parameters, but Dixon and Coles recognize that this static model needs several adjustments to account for real-world complexities such as fluctuating team performance.

One of the most significant extensions Dixon and Coles make to Maher's model is the introduction of time-varying parameters. In their framework, a team's attack and defense abilities are not constant throughout the season. Instead, recent match performances are considered more relevant by giving them more weight than older results, reflecting that a team's form may change over time. This is particularly true in football, where injuries, lineups changes, and managerial decisions can cause significant shifts in team performance. To do that, they formalize this idea by using a pseudo-likelihood function that emphasizes recent results over older ones. The weighting is based on an exponential decay function, ensuring that more recent matches have a greater impact on the estimated attack and defense parameters. This dynamic aspect improves the model's predictive power and adaptability.

Then, like Maher, Dixon and Coles account for the well-documented phenomenon of home-field advantage in football. They do so by introducing a parameter λ that adjusts the mean number of goals scored by the home team. This parameter reflects the fact that teams playing at home tend to perform better due to familiarity with the pitch, crowd support, and other factors.

Another critical enhancement introduced by Dixon and Coles is their treatment of score dependence. In Maher's original model, home and away scores are treated as independent. However, Dixon and Coles recognize that in many matches, particularly low-scoring ones, the scores are not entirely independent.
For example, if a team is trailing by one goal, they may take more risks and, in doing so, concede another goal or, conversely, score one themselves. To solve this, they propose a bivariate Poisson model that introduces a correlation between the scores of the two teams. This correlation is captured by a dependence parameter, ρ, which adjusts the probabilities of low-scoring outcomes. When $\rho = 0$, the model reduces to Maher's independent Poisson model. For most football matches, Dixon and Coles find that introducing this dependence parameter improves the model's accuracy, especially for predicting draws or closely contested games.

One of the main motivations for Dixon and Coles' work was to apply their model to betting markets, specifically the fixed-odds football betting market in the UK. By comparing the probabilities estimated by their model with those implied by the bookmakers' odds, they developed a strategy to identify "good bets" outcomes where the model's estimated probability was higher than the probability implied by the bookmakers' odds. To do this, they use a maximum likelihood approach to estimate the parameters for each team and match outcome. Their model is then used to predict the probabilities of home wins, draws, and away wins. These probabilities are compared to the bookmakers' odds, and a betting strategy is formulated based on these discrepancies.

The complexity of Dixon and Coles' model, particularly with the introduction of time-dependence and the bivariate Poisson structure, makes it impossible not to rely on numerical methods to compute maximum likelihood estimates of the model's parameters. The high dimensionality of the model—resulting from separate attack and defense parameters for each team and the home advantage parameter—requires sophisticated computational techniques, but they demonstrate that these calculations are manageable.

Dixon and Coles apply their model to historical data from the English football league, covering matches from the 1992-1995 seasons. By testing their model on this dataset, they show that it provides a good fit to observed match outcomes, particularly when it comes to predicting draws and low-scoring games, where the original independent Poisson model would struggle. Furthermore, they test their betting strategy using bookmakers' odds from the 1995-1996 season. By comparing the model's predictions to the odds offered by bookmakers, they demonstrate that their strategy produces a positive return, even when accounting for the inherent bias in bookmakers' odds. The introduction of dynamic parameters and score dependence significantly enhances the model's accuracy and profitability in the betting market.

Following Dixon and Coles' influential work, subsequent studies in football match prediction introduced the concept of rating systems, which provide a quantitative measure of a team's strength relative to its opponent. These systems offer a refined way of analyzing past performance data, with the goal of predicting future match outcomes and optimizing betting strategies. Unlike earlier models that primarily focused on raw scores or isolated match results, rating systems lead to more complex metrics to evaluate team superiority and translate it into actionable insights for betting markets.

1.3.3 Rating Systems

One such approach, detailed in the document Rating Systems for Fixed Odds Football Match Prediction (2003) [9], focuses on the Goal Superiority Rating System. This system assesses the difference in goal-scoring performance between two teams over their most recent matches, typically the last 4-6 games. The superiority of one team over another is computed as a match rating, which is calculated by subtracting the away team's goal superiority rating from that of the home team. For example, if Tottenham has a goal superiority rating of -3 (having scored 6 goals and conceded 9), and Leeds also has a rating of -3 (with 8 goals scored and 11 conceded), the match rating is 0, indicating that the game is likely to be evenly balanced.
This rating system draws upon a large dataset of historical football matches from the English leagues between 1993 and 2001, allowing the calculation of probabilities for match outcomes such as home wins, draws, and away wins. For instance, with a match rating of 0, the data suggests a 46% probability of a home win, a 28% chance of a draw, and a 26% likelihood of an away win. Higher ratings generally favor the home team, while lower ratings suggest a better chance for the away team. These probabilities were built upon a robust dataset of 16,272 matches, providing a solid foundation for predicting outcomes.

Figure 1.1 shows the relationship between match rating and the corresponding outcomes of football matches. The match rating quantifies the strength of the home team in respect to the away team based on recent goal-scoring performance, with higher ratings favoring the home team. The table provides data for home wins, draws, and away wins across a range of match ratings, from -10 (favoring the away team) to +10 (favoring the home team). As the match rating increases (i.e., the home team is expected to perform better), the percentage of home wins rises. For example, at a rating of 0 (neutral), the home win percentage is 45.6%, while at a rating of +10, it is 55.8%. Conversely, lower match ratings (favoring the away team) show a higher percentage of away wins. For instance, at a rating of -10, the away win percentage is 33.8%.
Beyond predicting match results, the system also facilitates the calculation of fair odds for each possible outcome. By converting the estimated probabilities into odds, the system allows bettors to compare these fair odds with those offered by bookmakers.
For example, a probability of 46.47% for a home win translates into fair odds of 2.15. This comparison is crucial for identifying value bets, where the bookmaker's odds are higher than the calculated fair odds, suggesting a profitable betting opportunity.

The profitability of this rating system is demonstrated through its application to English football matches during the 2001/02 season. By narrowing the focus

Match rating	Number of home wins	Number of draws	Number of away wins	% of home wins	% of home draws	% of away wins
-10	84	94	91	31.2%	34.9%	33.8%
-9	123	91	112	37.7%	27.9%	34.4%
-8	171	113	124	41.9%	27.7%	30.4%
-7	190	121	170	39.5%	25.2%	35.3%
-6	242	202	191	38.1%	31.8%	30.1%
-5	279	212	197	40.6%	30.8%	28.6%
-4	293	219	215	40.3%	30.1%	29.6%
-3	374	246	229	44.1%	29.0%	27.0%
-2	372	233	214	45.4%	28.4%	26.1%
-1	375	251	222	44.2%	29.6%	26.2%
0	414	259	235	45.6%	28.5%	25.9%
1	412	243	212	47.5%	28.0%	24.5%
2	401	220	189	49.5%	27.2%	23.3%
3	395	224	175	49.7%	28.2%	22.0%
4	391	177	137	55.5%	25.1%	19.4%
5	297	180	102	51.3%	31.1%	17.6%
6	260	146	131	48.4%	27.2%	24.4%
7	236	98	83	56.6%	23.5%	19.9%
8	197	94	56	56.8%	27.1%	16.1%
9	158	86	32	57.2%	31.2%	11.6%
10	125	57	42	55.8%	25.4%	18.8%
Total	**6468**	**3932**	**3602**	**46.2%**	**28.1%**	**25.7%**

Figure 1.1: Goal supremacy match ratings and historical result percentages from the Rating Systems for Fixed Odds Football Match Prediction (2003)

to matches with ratings between -2 and +2, the system achieved a 10.13% yield, showing that by consistently identifying and betting on value opportunities, it is possible to generate profits over time. This result underscores the effectiveness of rating systems in refining football match predictions and enhancing betting strategies by leveraging discrepancies between calculated fair odds and bookmaker offerings.

1.3.4 ELO Ratings

To conclude the state-of-the-art section on football prediction models, the paper Using ELO Ratings for Match Result Prediction in Association Football (2010) by Lars Magnus Hvattum and Halvard Arntzen [10] presents a significant development by applying ELO ratings to predict football match outcomes. ELO ratings, originally designed for ranking chess players, have been adapted here as a robust framework for rating football teams and forecasting results. At its core, the ELO

model assigns a numerical rating to each team based on their past performances, updating the rating after each match. A team's ELO rating increases when it wins and decreases when it loses, with the size of the adjustment depending on the relative ratings of the two teams involved.
For example, if a lower-rated team beats a higher-rated team, the adjustment to their ratings is greater than if the higher-rated team had won as expected. This aspect of the model reflects the notion that unexpected results carry more information and should cause greater shifts in team ratings.
Building on this basic framework, Hvattum and Arntzen introduce a refined version of the ELO system that accounts for goal differences. This is significant because it acknowledges that not all wins or losses are equal in competitive sports like football. A narrow 1-0 victory might not reflect a dominant performance, whereas a 3-0 win likely does. To capture this nuance, the model includes a goal-difference weighting, which means that more decisive wins result in larger increases to the winning team's ELO rating, while narrow victories lead to smaller adjustments. This enhancement makes the ELO rating more sensitive to the quality of a team's performance, rather than merely the outcome of the match.
To translate these ELO ratings into match predictions, the authors employ an ordered logit regression model. This statistical method uses the difference between the home team's and the away team's ELO ratings as the main covariate to predict the likelihood of three possible outcomes: a home win, a draw, or an away win. The greater the ELO difference between the two teams, the more likely the higher-rated team is to win. By applying ordered logit regression, the model does not just predict match outcomes in a binary win/loss format but assigns probabilities to each possible outcome.
Furthermore, Hvattum and Arntzen rigorously test their ELO-based models against a range of benchmark methods to evaluate their effectiveness. These benchmark methods include simpler models that rely on historical frequencies of match outcomes (such as the proportion of home wins, draws, and away wins) and more sophisticated models like those based on bookmaker odds. The authors find that while the ELO-based models consistently outperform naive models (such as those that assume equal probabilities for all outcomes), they fall slightly short when compared to bookmaker odds. This is likely because bookmakers incorporate a wider range of variables, such as player injuries, tactical changes, or insider knowledge, which are not included in the ELO model.

To assess the performance of their models, the authors use both statistical and economic evaluation metrics. Statistically, they measure the accuracy of their predictions using loss functions, which calculate the difference between the predicted probabilities and the actual outcomes. Economically, they simulate betting scenarios where bets are placed on the predicted outcomes, and the profitability of

these bets is measured.
While the ELO-based models perform well in these evaluations, particularly in terms of statistical accuracy, they are less successful in consistently beating the market (i.e., bookmaker odds) in real betting scenarios. This suggests that while ELO ratings are a valuable tool for predicting match outcomes, they may not fully capture the complexity of factors influencing real-world football matches.

Lastly, the goal-based extension of the ELO model represents an important innovation in sports prediction. By weighting victories and losses according to the margin of victory, the model is better able to differentiate between dominant performances and closely contested matches. This makes the ELO ratings more reflective of a team's true strength, particularly in cases where teams consistently win or lose by significant margins. This added detail enhances the predictive power of the ELO-based model, especially when applied to forecasting match outcomes over time.

1.4 Comprehensive framework

In reviewing the most influential models for football match predictions, each work contributes unique innovations that, when combined, create a clearer picture of how football score prediction has evolved. These innovations build upon one another, address the limitations of previous approaches and introduce new concepts that progressively enhance the accuracy and applicability of these models.
Maher's 1982 model was a pivotal starting point for statistical modelling in football. His application of the Poisson distribution provided a structured approach to modelling football scores. This framework was simple yet effective, particularly because it introduced the concept of team-specific parameters, something that had not been thoroughly explored before. This marked a significant shift away from earlier models that focused solely on random chance and led to the idea that team quality plays a decisive role in match outcomes.
Despite its novelty, Maher's model had clear limitations. The assumption that goals scored by the two teams were independent events overlooked the interactive nature of football, where one team's score often affects the tactics and behavior of the other team.

Additionally, Maher's model treated a team's attacking and defensive abilities as static over the course of a season, failing to account for the fluctuations in form that are common in football due to injuries, managerial changes, or other factors. This gap led to the next significant development: building on Maher's foundation, Dixon and Coles made two critical advancements.

First, they recognized the importance of incorporating time-dependence into their model. This dynamic aspect of the model ensured that teams on a winning streak or suffering from a downturn in form were more accurately represented, thus improving the model's predictive power.
Secondly, Dixon and Coles addressed the issue of score correlation. Unlike Maher's model, which assumed the goals scored by each team were independent, Dixon and Coles recognized that scores in football matches are often interdependent. To model this, they introduced a bivariate Poisson model that allowed for correlation between the scores of the two teams, reflecting the reality that one team's actions often influence the other's.
While Dixon and Coles focused on refining the Poisson framework to capture match dynamics more accurately, other researchers then turned to rating systems as an alternative approach. Unlike the Poisson-based models, which predict the exact number of goals scored, rating systems focus on providing a quantitative measure of team strength relative to the opponent. The simplicity of this system lies in its ability to condense a team's recent performance into a single, easily interpretable metric that can be used to predict the outcome of the match. While less sophisticated than the Poisson-based models, the rating system's broad application and ease of use made it a practical tool for those looking to apply football predictions to betting markets.
The final model reviewed is the ELO rating system: originally developed for ranking chess players, offers a more dynamic approach to rating teams based on their match outcomes. In the ELO framework, a team's rating increases when it wins and decreases when it loses, with the size of the adjustment depending on the relative strength of the opponent.

To sum up, these models represent key steps in the evolution of football predictions. The innovations they introduced, from team-specific parameters and dynamic weighting to rating systems and goal-difference adjustments, all contribute to the state-of-the-art methodologies used today. Each model has unique strengths, and together they provide the conceptual tools and insights necessary to develop more robust, accurate, and practical football prediction models, as it will be further explored in the next chapters of this thesis.

Chapter 2

Datasets and Pre-Processing

2.1 Forecasts to forecast

The central principle in this chapter is the idea of "forecasts to forecast". This concept is rooted in the understanding that prior forecasts—such as betting odds—already reflect a big amount of information that can be leveraged for future predictions. In the study "Using soccer forecasts to forecast soccer" (Wunderlich & Memmert) [11], it is demonstrated that betting odds outperform even mathematical models when it comes to predicting sports outcomes. Betting odds incorporate the collective knowledge of both bookmakers and the betting market, which includes public and private information on team form, injuries, and external factors. Therefore, betting odds are not raw predictions but an aggregate reflection of all available information at the time of the match, making them a powerful tool for forecasting.
The fundamental concept behind prediction markets (PMs) and the efficient market hypothesis (EMH) plays a crucial role here. As suggested by Friedrich Hayek (1945), a renowned Austrian-British economist and philosopher, and Eugene Fama (1970) [12], an influential American economist and proponent of the Efficient Market Hypothesis, markets solve information problems by efficiently aggregating dispersed information through prices. In competitive markets, prices (like those reflected in betting odds) represent the collective judgment of market participants, making them an effective predictor. This idea supports the efficiency of markets as mechanisms for gathering all available information, both public and private, and integrating it into a single forecast. Thus, betting odds can be seen as embodying a consensus of expert opinion that adjusts dynamically as new information emerges, aligning closely with the efficient market hypothesis.
As explained by Spann and Skiera (2008) [13], prices in competitive markets reflect the aggregation of dispersed information from various sources, thus providing an

aggregated view of expectations about future events; in this case, sports matches. This makes odds an ideal data source for predictive modeling, as they encapsulate far more information than raw match outcomes.
Building on this idea, the dataset used in this thesis is based on historical betting odds from various bookmakers, providing a comprehensive look at the dynamics of betting markets. These odds are collected for a wide range of soccer matches, spanning multiple leagues and seasons. The dataset includes both opening and closing odds for home wins, draws, and away wins, capturing how the betting market's expectations change over time as more information becomes available. This dual inclusion allows for a deeper exploration of market behavior, as opening odds reflect early expectations, while closing odds show the market's final consensus right before the match.

To build the dataset, a custom web scraping tool was developed using Python, leveraging libraries such as Playwright and BeautifulSoup. These tools were chosen for their ability to interact with dynamic websites and efficiently extract relevant data. The script automates the collection of odds from major public betting sites, gathering not only the final odds but also the evolution of those odds over time.
The scraping process involved navigating to specific match pages on betting platforms, extracting the relevant data from these pages: for each match, the scraper collected data on odds provided by several major bookmakers, including Pinnacle, William Hill, 1xBet, and Bet365, ensuring a broad representation of market sentiment. Additionally, to capture the full scope of the betting market, the tool was programmed to scrape both 1X2 odds (for home win, draw, or away win) and over/under odds for goal totals. This allows for a detailed analysis of how different types of bets evolve in response to market signals. The data is organized by league, match, and bookmaker, ensuring consistency and comparability across different leagues and seasons.

2.2 Implied Probability

The odds we obtain through scraping are raw data, reflecting the bookmakers' assessments of match outcomes. However, these odds need to be processed before they can be used for analysis. Raw odds are typically influenced by various factors, including bookmakers' margins, which are built in to ensure profitability. As a result, the sum of the implied probabilities derived from these odds often exceeds 100%, meaning the raw odds cannot directly serve as true indicators of the match probabilities. To standardize the data and make it suitable for predictive modeling, we perform a conversion step: transforming odds into implied probabilities.

To explain implied probabilities and how they are extracted from betting odds, we can refer to the principles discussed in the paper "A Better Way to Normalize Odds to Probability" by Erik Štrumbelj [14], which discusses different methods of converting odds into probabilities. When bookmakers set odds for a match, they don't just reflect pure probabilities; instead, they factor in a margin to ensure profit. This is why the sum of the implied probabilities, derived directly from the odds, is typically greater than 100%. To derive true probabilities from these odds, we need to adjust or normalize the odds.
Two most used methods for this are basic normalization and Shin's model [14].

2.2.1 Basic Normalization

The simplest approach to convert odds into probabilities is to use basic normalization. If you have decimal odds o_i or each possible outcome, the implied probability for each outcome is calculated as the inverse of the odds:

$$\pi_i = \frac{1}{o_i}$$

However, the sum of these implied probabilities usually exceeds 1 due to the bookmaker's margin. To adjust for this, you can normalize the probabilities by dividing each implied probability by the sum of all implied probabilities:

$$p_i = \frac{\pi_i}{\sum_{i=1}^{n} \pi_i}$$

This method ensures that the probabilities sum to 1, providing a straightforward way of extracting probabilities from odds.

2.2.2 Shin's Model

A more sophisticated approach to calculating implied probabilities is Shin's model, which assumes that some bettors have insider information and that bookmakers adjust their odds to account for this. Shin's model adjusts the implied probabilities by estimating the proportion of bets placed by informed traders. The model adjusts for the fact that some participants may have more accurate information than others, leading to a slight distortion in the odds.

The Shin probability p_i for each outcome is calculated as:

$$p_i = \frac{z + (1-z)\pi_i^2}{\sum_{j=1}^{n} z + (1-z)\pi_j^2}$$

Where:

- p_i is the adjusted probability for outcome i,
- z represents the proportion of bets from informed bettors (insiders),
- $1 - z$ represents the proportion of bets from uninformed bettors,
- π_i is the implied probability for outcome i, based on the bookmaker's odds,
- $\sum_{j=1}^{n}$ is the sum over all possible outcomes.

To estimate the value of z, we can use the condition that the sum of all adjusted probabilities p_i must equal 1. This leads to the iterative calculation of z, with an initial guess of $z_0 = 0$.

Having established the importance of betting odds as a key source for forecasting match outcomes, the next critical challenge is to determine which bookmakers provide the most accurate and favorable odds. Not all bookmakers set their odds in the same way; different factors such as market share, risk tolerance, and target audience can influence the odds they offer. The accuracy of implied probabilities is highly dependent on the quality of the raw odds provided. Bookmakers with larger market shares and more sophisticated risk management systems, for example, tend to offer odds that reflect aggregate market sentiment more closely. Meanwhile, smaller or niche bookmakers may offer odds that deviate significantly from the consensus due to local biases or risk strategies, potentially leading to less reliable probability estimates.
This evaluation involves comparing the odds provided by different bookmakers, not just in terms of their competitiveness but also their predictive power over time. By systematically analyzing odds from multiple bookmakers and their alignment with actual match outcomes, we can identify which bookmakers offer the most accurate reflection of the true probabilities.
The step is critical because it will directly influence the reliability of the predictive models developed in this thesis. If certain bookmakers are consistently offering odds that lead to more accurate implied probabilities, these bookmakers can be prioritized in the modeling process to enhance the overall predictive performance.

2.3 Evaluating predictions

In order to assess the accuracy of the implied probabilities provided by different bookmakers, we need a robust and valid method for evaluating predictions. This involves applying specific scoring rules that measure how well the probabilities reflect the actual outcomes.
The three most commonly used methods for this evaluation are the Ranked Probability Score (RPS), the Brier Score, and the Ignorance Score [15].

2.3.1 Ranked Probability Score

The Ranked Probability Score (RPS) is a proper scoring rule used to evaluate probabilistic forecasts where there are multiple, ordered outcomes. It is particularly useful for assessing predictions in scenarios where the outcomes have a natural ordering, such as in football matches with three possible outcomes: home win, draw, or away win.
The formula for RPS is as follows:

$$RPS = \sum_{i=1}^{r-1} \sum_{j=1}^{i} (p_j - o_j)^2$$

Where:

- p_j is the predicted cumulative probability for outcome j,
- o_j is the actual cumulative outcome (which is 1 for the correct category and 0 for all others),
- r is the number of possible outcomes (for football: 3 outcomes—home win, draw, away win).

Main characteristics of the RPS:

- **The cumulative probability** means the probabilities for earlier outcomes are added together before being squared.
- **Distance-sensitive:** The RPS penalizes predictions more heavily when the forecasted probabilities for incorrect outcomes are further away from the correct result. For example, if the outcome is a home win, the score will be lower if the model assigned a high probability to a draw than if it assigned a high probability to an away win.

- **Non-local:** RPS is influenced not only by the probability assigned to the actual outcome but also by the probabilities assigned to the incorrect outcomes, considering the entire probability distribution.

2.3.2 Brier Score

The Brier Score is another famous scoring rule used to evaluate the accuracy of probabilistic predictions. It measures the squared difference between the predicted probabilities and the actual outcome for each possible outcome. Unlike RPS, the Brier Score does not take into account the ordering of outcomes.

The formula for Brier Score is:

$$\text{Brier Score} = \sum_{i=1}^{r} (p_i - o_i)^2$$

Where:

- p_i is the predicted probability for outcome i,
- o_i is 1 for the actual outcome and 0 for all others,
- r is the number of possible outcomes (for football: 3 outcomes—home win, draw, away win).

Unlike the concept of locality RPS had, the Brier Score evaluates each outcome independently, considering all deviations from the actual outcome equally, regardless of how far the incorrect outcomes are from the true result. A prediction assigning high probability to a home win will be penalized just as much whether the actual result is a draw or an away win, even though a draw may be considered “closer” to a home win. It’s more versatile and easier to interpret than the RPS, that’s why it’s widely used in practice.

2.3.3 Ignorance Score

The Ignorance Score (or log score) is a local scoring rule that focuses entirely on the probability assigned to the actual outcome, ignoring the probabilities assigned to other possible outcomes. It penalizes forecasts that assign low probability to the correct outcome, with higher penalties for extremely low probabilities.

Formula:

$$\text{Ignorance Score} = -\log_2(p(Y))$$

Where:

- $p(Y)$ is the probability assigned to the actual outcome.

Main features of Ignorance Score

- **Locality and sensitiveness to distance:** The Ignorance Score only evaluates the probability assigned to the actual outcome and disregards the rest of the probability distribution. It totally ignores the probabilities proximity to the actual result.

- **Exponential penalties:** Forecasts that assign very low probabilities to the correct outcome are heavily penalized. For example, predicting a very unlikely outcome with high certainty will result in a very high Ignorance Score if the outcome doesn't occur.

The Ignorance Score is often chosen when the goal is to maximize the probability assigned to the correct outcome, without worrying about how well the forecast handles the incorrect outcomes. In many models, the goal is to assign probabilities to just one outcome, making the Ignorance Score a suitable choice for evaluation.

2.4 Evaluation results

The matches analyzed span from the 2017 to 2022 seasons, across multiple top football leagues, including:

- Premier League (England): 2280 matches,
- Ligue 1 (France): 2189 matches,
- La Liga (Spain): 2279 matches,
- Serie A (Italy): 2279 matches,
- Serie B (Italy): 1863 matches,
- Serie C (Italy): 7033 matches.

Odds Provider	RPS	RPS(Shin)	Ign. Score	Brier Score
England PL				
Pinna Opening	0.198	1.059	1.385	0.568
Pinna Closing	**0.194**	1.059	**1.370**	**0.562**
Bet365 Opening	0.197	1.059	1.382	0.567
WH Opening	0.197	1.059	1.385	0.568
UNOX Opening	0.197	1.059	1.384	0.568
France Ligue 1				
Pinna Opening	0.202	1.059	1.434	0.593
Pinna Closing	**0.199**	1.059	**1.419**	**0.586**
Bet365 Opening	0.202	1.059	1.431	0.592
WH Opening	0.202	1.059	1.433	0.593
UNOX Opening	0.203	1.059	1.438	0.595
Spain La Liga				
Pinna Opening	0.198	1.083	1.427	0.590
Pinna Closing	**0.195**	1.083	**1.413**	**0.583**
Bet365 Opening	0.197	1.083	1.424	0.588
WH Opening	0.198	1.083	1.424	0.589
UNOX Opening	0.199	1.083	1.432	0.592
Italy Serie A				
Pinna Opening	0.191	1.042	1.380	0.566
Pinna Closing	**0.188**	1.042	**1.363**	**0.559**
Bet365 Opening	0.191	1.042	1.377	0.565
WH Opening	0.191	1.042	1.378	0.565
UNOX Opening	0.192	1.042	1.383	0.568
Italy Serie B				
Pinna Opening	0.208	1.075	1.525	0.640
Pinna Closing	**0.206**	1.075	**1.517**	**0.635**
Bet365 Opening	0.209	1.075	1.529	0.641
WH Opening	0.209	1.075	1.530	0.641
UNOX Opening	0.210	1.075	1.533	0.643
Italy Serie C				
Pinna Opening	0.207	1.060	1.511	0.632
Pinna Closing	**0.207**	1.060	**1.508**	**0.630**
Bet365 Opening	0.209	1.060	1.520	0.636
WH Opening	0.209	1.060	1.520	0.636
UNOX Opening	0.209	1.060	1.520	0.636

Table 2.1: RPS, Ignorance Score, and Brier Score for Different Odds Providers Across Leagues

The table presents the performance of different bookmakers' odds (Pinnacle, Bet365, William Hill, and UNOX) evaluated using the Ranked Probability Score (RPS), Shin-normalized RPS, Ignorance Score, and Brier Score. These scores test whether the results would change based on the different evaluation techniques. Overall, the results remain relatively consistent across the different metrics, although some key insights emerge.

One of the most significant observations is that Shin-normalized RPS produces less favorable results across all cases, compared to the standard RPS. This is because the Shin method is optimal for evaluating individual matches, where it adjusts probabilities based on the assumption of insider information. However, when applied to a batch of matches (as in our case), this adjustment leads to a flattening of probabilities due to the underlying assumption of insider knowledge being present across all matches. This effect diminishes the differentiation between probabilities for each outcome, which is why the Shin-normalized scores are consistently higher and less reflective of actual performance.
Therefore, we conclude that Shin normalization is not suitable for batch processing in this context and does not provide useful insights for evaluating large datasets of matches.

Across all leagues and metrics, the Pinnacle closing odds consistently show the best performance. This can be attributed to Pinnacle's reputation as one of the world's leading bookmakers, known for employing some of the best traders in the industry. Pinnacle is renowned for offering higher odds compared to competitors because they possess superior information and insights, allowing them to set more accurate probabilities. While offering higher odds results in a lower margin (commonly known as the "vig"), Pinnacle compensates for this by dominating the market with more competitive prices, attracting larger betting volumes. Their strategy of prioritizing accuracy and market competitiveness makes their closing odds particularly reliable, as they reflect the most up-to-date and comprehensive information at the moment the market closes. This explains why their odds consistently outperform other bookmakers in terms of prediction accuracy.

All these observations reinforce the idea that closing odds, the final odds set just before the match starts, are more reliable for modeling match outcomes. These odds reflect all the available information, including the latest insights and market dynamics, which have been incorporated in the final moments before the market closes. According to Fama's Efficient Market Hypothesis, market prices (or odds, in this case) reflect all available information, and this is why Pinnacle's closing odds are particularly accurate. When comparing the performance of Bet365, William Hill, and UNOX, we observe similar results across these bookmakers. They show

small variations but are generally consistent in performance.
For example, in Serie A, the RPS for Bet365 opening odds is 0.191, while William Hill and UNOX show almost identical scores. These results suggest that while these bookmakers provide relatively accurate odds, their performance does not surpass Pinnacle's, particularly when it comes to closing odds.

As such, we will use Pinnacle's closing odds as the primary data source to train our predictive model in the following chapters, leveraging their accuracy to improve the model's overall performance.

2.5 Dataset Structure

The dataset we are presenting here serves as the backbone for the model we will develop and analyze in subsequent chapters. To ensure accuracy and relevance in our analysis, we have generated separate datasets for each league—Premier League, Ligue 1, La Liga, Serie A, Serie B, and Serie C. This distinction is important because each league has its own unique characteristics that we will explore in detail, and keeping them separate prevents any confusion among leagues. Each dataset will be tailored to capture the specific dynamics of the individual competition, allowing for more precise modeling and predictions.

The dataset includes several key parameters that will be used to build and refine the model:

- **home_team, away_team**: the names of the teams playing at home and away, respectively.
- **home_score, away_score**: the number of goals scored by the home and away teams in the match.
- **odd_home_avg, odd_draw_avg, odd_away_avg**: the average odds across multiple bookmakers for a home win, draw, and away win, respectively.
- **date**: the date when the match was played.
- **home_pinna_opening, draw_pinna_opening, away_pinna_opening**: the opening odds for a home win, draw, and away win from Pinnacle.
- **home_pinna_closing, draw_pinna_closing, away_pinna_closing**: the closing odds for a home win, draw, and away win from Pinnacle.
- **home_wh_opening, draw_wh_opening, away_wh_opening**: the opening odds for a home win, draw, and away win from William Hill.

- **home_bet365_opening, draw_bet365_opening, away_bet365_opening**: the opening odds for a home win, draw, and away win from Bet365.
- **home_unox_opening, draw_unox_opening, away_unox_opening**: the opening odds for a home win, draw, and away win from UNOX.
- **pinna_opening_uo_over, pinna_opening_uo_under**: Pinnacle's opening odds for the match total goals being over or under a specific value.
- **result**: the outcome of the match, typically coded as home win, draw, or away win.

Chapter 3

Model

Having established a robust dataset and applied various evaluation metrics to assess bookmaker odds, we now turn our attention to the development of the predictive model. This chapter will explain how the data will be leveraged to create a model that accurately forecasts match outcomes based on historical odds, match results, and other key features.
The model will take into account the dynamic nature of betting markets, using the implied probabilities and closing odds—as discussed in the previous chapters—as the foundation for building a reliable prediction system. Additionally, we'll explore how the league-specific datasets will allow us to capture the nuances and characteristics unique to each competition, improving the model's overall accuracy and adaptability.

We will also discuss the methodologies and techniques employed in building the model, including machine learning algorithms, parameters optimizations, feature selection, and training strategies, ensuring that the model is capable of handling the complexities of football match prediction.
In this next chapter, the focus will shift to how the parameters extracted and processed in the dataset play an integral role in shaping the model and how different variables interact to produce accurate predictions.

3.1 Core Methods and Functions

In this section, we will explore the core methods and functions that will be used repeatedly throughout the model development. These methods include the Poisson distribution, optimization using minimization, and the Sum of Squared Errors (SSE), all of which play a critical role in enhancing the accuracy and reliability of match outcome predictions.

To better understand the core methods and functions, some Python code snippets are provided to illustrate their implementation and usage in the model development process.

3.1.1 Refined Poisson Approach

The Poisson distribution is a core element of our predictive model, particularly for forecasting the number of goals scored by each team in a football match.
The distribution [3] is based on the assumption that goals in football are independent events that occur at a fixed rate, making it a suitable tool for predicting outcomes. However, as already discussed in chapter 1, raw Poisson predictions often require adjustments (inflation) to better fit real-world results.

To predict the probabilities of different match outcomes—home win, draw, and away win—we compute a Poisson matrix that estimates the likelihood of a specific goal count for both teams in a match. Each cell of the Poisson matrix represents the probability of the home team (A) scoring exactly x goals and the probability of the away team (B) scoring exactly y goals.
For example, the probability of team A scoring 2 goals and team B scoring 1 goal is computed using the Poisson probability formula:

$$P(\text{team A scores } x \text{ goals}) = \frac{\lambda_A^x \cdot e^{-\lambda_A}}{x!}$$

where λ_A is the expected goal rate for team A.

Similarly, for team B scoring y goals:

$$P(\text{team B scores } y \text{ goals}) = \frac{\lambda_B^y \cdot e^{-\lambda_B}}{y!}$$

where λ_B is the expected goal rate for team B.

Poisson Matrix

By multiplying these probabilities for each combination of x (goals by the home team) and y (goals by the away team), we populate the Poisson matrix. The image below shows an example of such a Poisson matrix, where the values represent the probability of each specific outcome. The rows represent the number of goals scored by the away team, while the columns represent the number of goals scored by the home team. For instance, the probability of the home team scoring 2 goals and the away team scoring 1 goal (a 2-1 victory for the home team) is given by the intersection of row 1 and column 2.

					Home			
		0	**1**	**2**	**3**	**4**	**5**	**6**
Away	**0**	0,0202	0,0425	0,0446	0,0312	0,0164	0,0069	0,0024
	1	0,0364	0,0765	0,0803	0,0562	0,0295	0,0124	0,0043
	2	0,0328	0,0689	0,0723	0,0506	0,0266	0,0112	0,0039
	3	0,0197	0,0413	0,0434	0,0304	0,0159	0,0067	0,0023
	4	0,0089	0,0186	0,0195	0,0137	0,0072	0,0030	0,0011
	5	0,0032	0,0067	0,0070	0,0049	0,0026	0,0011	0,0004
	6	0,0010	0,0020	0,0021	0,0015	0,0008	0,0003	0,0001

lambdaH	2,100	**1**	44,98%	2,223
lambdaA	1,800	**X**	20,78%	4,812
		2	33,52%	2,984

Figure 3.1: Example Poisson Matrix: Probabilities of Various Score Outcomes for Home and Away Teams.

Once the matrix is computed, the next step is to sum the probabilities of the relevant results to calculate the probabilities of a home win, draw, and away win:

- **Home win**: The probability of a home win is the sum of all cells where team A scores more goals than team B, i.e., $P(\text{Home win}) = \sum_{x>y} P(x, y)$.
- **Draw**: The probability of a draw is the sum of all cells where team A's goals equal team B's goals, i.e., $P(\text{Draw}) = \sum_{x=y} P(x, y)$.
- **Away win**: The probability of an away win is the sum of all cells where team B scores more goals than team A, i.e., $P(\text{Away win}) = \sum_{x<y} P(x, y)$.

This way, by analyzing the full Poisson matrix, we can compute the overall likelihood of each match outcome.

Inflaction

While the Poisson distribution provides a useful framework for modeling football scores, raw Poisson models are often not ideal in practice. As described in chapter one, certain match results—such as high-scoring games or upsets—are not well captured by the basic Poisson model.
To address this, we introduce the concept of matrix inflation. This involves adjusting the probabilities in specific parts of the matrix to account for known biases in the Poisson model:

- **Low-probability results**, such as very high scores or unlikely outcomes, tend to be underestimated by the raw Poisson model. To correct for this, we can inflate the probabilities in those cells.
- **High-probability results** may need to be slightly deflated to reflect real-world data, expecially some leagues.

The inflation is applied using an adjustment factor, based on historical data, that boosts or reduces the probability of specific cells. This allows the model to capture rare events more accurately, while still relying on the Poisson distribution for the majority of predictions.

3.1.2 Sum Squared Error

The Sum of Squared Errors (SSE) is a widely-used metric for measuring the difference between observed and predicted values. In the context of our model, SSE compares the implied probabilities derived from bookmaker odds with the predicted match outcomes. It is computed by taking the difference between each observed value (the actual outcome) and the corresponding predicted value, squaring these differences, and then summing them across all observations. The formula for SSE is:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where:

- y_i represents the observed value (e.g., the actual result of the match),
- $\hat{y}_i$ represents the predicted value (e.g., the predicted probability of a match outcome),
- n is the total number of observations.

The SSE highlights the discrepancy between the predicted results and the actual outcomes, with larger values indicating a greater difference and smaller values indicating a closer match.
By minimizing the SSE, we can fine-tune the model to more accurately reflect real-world data, improving its predictive performance. This method is crucial for assessing the overall accuracy of the model and guiding optimizations.

3.1.3 Minimization Function

Minimization is a mathematical process used to optimize certain parameters by reducing the difference between observed and predicted values, as will be explained later in the context of predicting match outcomes or calculating dominance scores. To do so, the objective of minimization is to adjust parameters in a way that the error or cost function—representing the discrepancy between predictions and actual outcomes—is minimized. This process plays a key role in the model, as it is applied multiple times to optimize different aspects to fine-tune certain parameters.

The core of minimization lies on the objective function: this is a mathematical function representing the error or difference between observed and predicted data points. In our case, we might minimize the difference between predicted match probabilities (using the Poisson model or odds-based calculations) and actual outcomes. The goal is to find parameter values that reduce this difference as much as possible.

In this thesis, minimize function from the SciPy Python library is used to handle this optimization.

```python
from scipy.optimize import minimize

# General form of minimize function
result = minimize(
    fun=objective_function,
    x0=initial_guess,
    args=(),
    method=optimization_method,
    tol=tolerance_level
)
```

In this code:

- fun: The objective function to be minimized. This function calculates the difference or error we want to minimize.
- `x0`: The initial guess for the parameters we are optimizing. This serves as the starting point for the minimization process.
- `args`: Extra arguments required by the objective function. For example, this can include observed probabilities, odds, or other parameters.
- `method`: The optimization method to use. Common methods include:
  - Powell: A derivative-free optimization method
  - BFGS: A quasi-Newton method for unconstrained problems.
- `tol`: Optimization methods use convergence criteria to decide when to stop iterating. These can be based on function value tolerance, gradient norm tolerance, or step size tolerance.

Minimize method Choice

- `Powell's method` performs a series of line searches along different directions without using gradients. While it can handle complex, non-convex problems, it is generally slower than gradient-based methods and not recommended for high-dimensional or large-scale problems. It's often a good fallback when derivative information is unavailable or unreliable.
- `The Nelder-Mead` algorithm is a derivative-free method that uses the geometry of a simplex to explore the function space, making it effective for problems where gradients are hard or impossible to compute. It's particularly useful for small-scale problems or when the objective function is noisy or non-differentiable. However, its performance can degrade in high-dimensional spaces, and it may converge slowly or get stuck in local minima on complex, non-convex problems.
- `L-BFGS-B` is a quasi-Newton method that approximates the Hessian matrix (second-order derivatives) using limited memory, making it efficient for high-dimensional problems. It leverages gradient information to rapidly converge, particularly on smooth, convex problems. A key advantage is that it supports box constraints, allowing variable bounds without altering the core algorithm. It's well-suited for large-scale problems, but since it relies on gradient information, it's not ideal for non-smooth or noisy objective functions.

3.2 Model Architecture

This section will deeply describe the model built, illustrating how it integrates key concepts from previous works—particularly those discussed in Chapter 1—into a cohesive framework for predicting match outcomes.
The model leverages the Poisson distribution, rating systems, and dominance to form a comprehensive predictive system. These methods are tied together with optimization techniques, notably the minimize function, which ensures that the predicted probabilities align with real-world outcomes as accurately as possible.

3.2.1 Main ideas

This model builds upon the most promising and important features derived from prior models: the key concept revolves around leveraging odds to calculate probability distributions, with several enhancements inspired by past works.

First and foremost, the model is grounded in the principles of Eugene Fama's (1970) Efficient Market Hypothesis (EMH) [12], which posits that market prices (in this case, betting odds) reflect all available information. Fama's insight suggests that odds are sufficient to build a reliable dataset because they incorporate the collective knowledge of market participants, both public and private. As such, this model exploits the historical odds as the most valuable source of information, forming the backbone of our predictive approach.

Following the work of Dixon and Coles [8], the model acknowledges the importance of time in weighing past odds. Recent odds are given more weight, while older odds are progressively de-emphasized, ensuring that only relevant, up-to-date information influences the model's predictions. This is a direct adaptation of the time-weighted approach Dixon and Coles applied to goals scored in previous matches. In this case, the focus is on the evolving market perception captured through betting odds, rather than just goals.

A new parameter introduced in this model is dominance. Unlike previous models that separated a team's attacking and defensive strengths, the dominance parameter measures the relative strength of the home team compared to the away team. It represents how much stronger (or weaker) the home team is, serving as a more holistic measure of superiority. This parameter is not a stand-alone measure of strength but a comparative one, providing insight into the specific matchup between two teams.

Additionally, team ratings, which were explored in various rating models discussed in Chapter 1, are central to the model. For each league, a distinct rating table is computed to reflect the relative strength of each team. This ensures that the model can account for the specific characteristics of different leagues, providing more accurate predictions based on the competition unique dynamics.

Lastly, like the ELO rating system, which adjusts over time, the model incorporates a sliding window that considers only recent historical odds. By adjusting the weight of older data, the model remains responsive to changes in team form and market conditions, ensuring that predictions remain relevant and accurate over time.

3.2.2 Dominance

The model begins by calculating the dominance for each match, a key parameter introduced in this model that reflects the home team's superiority or inferiority relative to the away team. The dominance can be either positive (indicating a stronger home team) or negative (indicating a weaker home team compared to the away team).
The model is designed for each individual league, using the historical odds for that league. To compute dominance, the model takes past odds from the dataset, specifically using Pinnacle closing odds, as demonstrated in Chapter 2, since these odds are known to provide the most reliable insights into match outcomes.

The raw odds extracted from the dataset need to be normalized. In this model, standard normalization is used to convert the odds into probabilities, as opposed to Shin normalization. As explained in Chapter 2, this choice is made because standard normalization provides more promising results when applied to batches of matches, rather than focusing on individual games with potential insider information.

Implied Total Goals

A new parameter introduced in the model is implied total goals (TG). This value represents the total number of goals expected in a match, as inferred from the normalized betting odds. The calculation of implied total goals helps the model estimate the likelihood of specific match outcomes, by predicting the total number of goals that might be scored. To optimize this process and improve computation speed in the following, it is precomputed the implied total goals and stored them in a Python dictionary called `z_poisson`.
The precomputation of implied total goals is necessary because raw Poisson distributions are capable of calculating probabilities only for integer values of goals.

However, in practice, the total number of goals expected in a match is given from the odds as a decimal value, such as 2.5 goals. To address this, it is used an adapted version of the Poisson distribution that can work with non-integer values to approximate the likelihood of scoring fractional goals.

TG Probability	**Implied total goals**
0.31716	3.5
0.36956	3.25
0.42319	3.0
0.48145	2.75
0.54381	2.5
0.60933	2.25
0.67667	2.0
0.74396	1.75
0.81884	1.5

Table 3.1: Extract some value from z_poisson dictionary

In the table provided an extract from the z_poisson dictionary is present: the left column shows the probability that more than 2.5 goals will be scored (TG probability), while the right column shows the implied total goals (the expected number of goals in the match) that correspond to these probabilities. Doing this precomputation, the vig is removed and the model is able to extract the expected number of goals from betting odds.

Dominance Computation

Two primary functions are involved in this computation: `minimize_dominance` and `objective_function_dominance`. These functions work together to optimize the dominance parameter, ensuring that the model's predicted probabilities align with observed data from the betting markets.
The first function, `objective_function_dominance`, calculates the difference between the predicted home win probability (using the Poisson adapted model) and the observed home win probability derived from normalized betting odds (from the historical league dataset). The goal of the minimization process is to adjust the dominance parameter to minimize this difference.

Listing 3.1: Objective Funtion Dominance

```
def objective_function_dominance(dominance: float,
    implied_tg: float, home_normalized_probability: float) ->
     float:
     lambda_home = (dominance + implied_tg) / 2
     lambda_away = implied_tg - lambda_home

     calculated_home = lmbda_to_probabilities(lambda_home,
    lambda_away, inflation=True)[0]

     difference = abs(calculated_home -
    home_normalized_probability)

     return difference
```

Parameters:

- `dominance`: The variable being optimized, representing the home team's relative strength.
- `implied_tg`: The total implied goals for the match, used to calculate the expected goals for both teams.
- `home_normalized_probability`: The observed probability of a home win, derived from normalized odds.

This function uses the Poisson distribution to compute the home win probability (calculated_home) based on the dominance and implied total goals. The function then returns the absolute difference between this computed value and the observed home win probability.

The second function, `minimize_dominance`, uses the Powell optimization method to find the dominance value that minimizes the difference calculated by the objective function.

Listing 3.2: Minimizing the dominance variable

```
def minimize_dominance(implied_TG: float,
  home_normalized_prob: float) -> float:
    minimized_dominance = minimize(
        objective_function_dominance,
        0.1,
        args=(implied_TG, home_normalized_prob),
        method="Powell",
        tol=1e-6)
    return minimized_dominance.x[0]
```

List of Parameters:

- `implied_TG`: The total implied goals for the match, which is used to compute expected goals for both teams.
- `home_normalized_prob`: The normalized probability of the home team winning, derived from the betting odds.

This function calls `objective_function_dominance` and uses it to minimize the difference between the calculated and observed probabilities. The Powell method is used to iteratively adjust the dominance value until the difference is minimized, yielding the optimized dominance parameter.

3.2.3 Ratings

The model incorporates a dynamic rating system to represent the relative strength of each team. Ratings play a critical role in predicting match outcomes by quantifying the performance of teams based on historical data. Each team's rating evolves over time as new matches are played, with adjustments based on the last several matches. These ratings are used to inform the model's Poisson-based probability estimates for outcomes.

The concept of team ratings is inspired by the ELO rating system, which is widely used in various competitive sports, as discussed in chapter 1, to adjust a team's strength based on match results. In this model, ratings are initially set to a baseline value, and they are dynamically updated after every match using the Sum of Squared Errors (SSE) to optimize the predictions.

Dynamic Ratings and the Sliding Window

The ratings for each team are computed using a sliding window approach, where only the last 5 to 7 match days are considered. This method ensures that the ratings reflect a team's most recent performance, removing older matches from the computation.
The idea of giving more importance to recent matches and removing older ones is inspired by Dixon and Coles [8], who similarly down-weighted older data, and ELO, which emphasizes the impact of recent matches in rating adjustments.
When a new match is played, the oldest match is removed from the batch, ensuring that the ratings remain current and adaptive. More recent matches carry greater weight in the calculations, ensuring that the ratings reflect the team's latest form. This process provides the model with up-to-date information and avoids over-reliance on outdated performance data.

Rating Computation

The `get_ratings` function in the model is responsible for retrieving the team ratings for each match, and adjusting the rating available in the model at the moment of the computation. These ratings are then fed into the Poisson model to estimate match probabilities.

The computation of team ratings is done using the minimize function, which optimizes the ratings to minimize the Sum of Squared Errors (SSE) between predicted and observed outcomes. The minimize function returns the optimized ratings in `result.x`, which represents the updated ratings for each team. The `result.fun` value is important, as it provides the minimized SSE and indicates the accuracy of the computation. A lower `result.fun` value means that the model has achieved better predictive accuracy.
The `get_ratings` function can be applied using a batch of 7, 6, or 5 championship days, depending on which batch provides the best result in terms of the `result.fun` value. This sliding window approach ensures that only the most recent match data is used for the rating computation, and the batch with the lowest `result.fun` is selected to achieve the most accurate ratings.

The function also computes the home advantage, a key parameter reflecting the trend that home teams statistically perform better than away teams. This home advantage is incorporated into the dominance calculation, where the home team's rating is adjusted upwards, and the away team's rating is adjusted downwards. The home advantage is further optimized as part of the SSE minimization process.

The table 3.2 is an example of computed ratings for Premier League teams, which are updated dynamically after every match. This table can be interpreted as a snapshot of the state of the ratings at a specific point in time, where the ratings of each team will evolve with new match data.

Sum of Squared Errors (SSE) Optimization

The `minimize_sse` function is responsible for optimizing the team ratings using the Sum of Squared Errors (SSE). This function iteratively adjusts the ratings to minimize the error between predicted outcomes and actual match results, ensuring that the model becomes more accurate over time. The Sum of Squared Errors (SSE) is calculated by comparing the predicted probabilities of the match outcomes (home win, draw, away win) with the actual outcomes. The ratings are then adjusted accordingly to reduce this difference.

Listing 3.3: Minimizing SSE for the difference between observed dominance and predicted dominance

```
def minimize_sse(values, reversed_pd_data, names_index,
   err_factor):
    result = minimize( minimize_sse,
                       initial_values,
                       args=( reversed_pd_data,
                              names_index,
                              err_factor),
                       method='Nelder-Mead',
                       tol=1e-5,
                       options={'maxfev': 100000})
    return result.x
```

Parameters:

- `values`: The initial values used to compute the ratings.
- `reversed_pd_data`: this variable contains the historical performance data for the teams, allowing the function to compare predicted outcomes to actual results.
- `names_index`: this parameter stores the index mapping of team names to their ratings.
- `err_factor`: A scaling factor used to control the weight of the error during optimization.

Team	**Rating**
Arsenal	0.82319
Aston Villa	0.26208
Bournemouth	-0.85756
Brentford	0.11612
Brighton	0.39320
Chelsea	0.19739
Crystal Palace	-0.28362
Everton	-0.38485
Fulham	-0.50177
Leeds	-0.41744
Leicester	-0.39466
Liverpool	0.92079
Manchester City	1.25639
Manchester Utd	0.65606
Newcastle	0.76501
Nottingham	-0.65587
Southampton	-0.99030
Tottenham	0.16318
West Ham	-0.44354
Wolves	-0.62380
home_factor	0.39375

Table 3.2: Ratings for Premier League Teams and Home Factor

The `minimize_sse` function uses the Nelder-Mead optimization method to minimize the error, adjusting the ratings until the difference between predicted and actual match outcomes is minimized. The optimized ratings are then used for future predictions, allowing the model to become progressively more accurate as more data is fed into it.

3.2.4 Odds Computations

Once all team ratings and parameters have been computed, the model can use them to calculate the odds for any possible match between rated teams. This is achieved through the function `calculate_single_match`, which combines various inputs, including team ratings and goal-scoring probabilities, to generate accurate odds for home win, draw, and away win outcomes.

The function `calculate_single_match` simulates a match between two teams, using the Poisson distribution to estimate the expected number of goals each team will score based on their ratings and the dominance factor. It then computes the probabilities for each possible match outcome (home win, draw, away win) and converts these probabilities into odds.

The `calculate_single_match` function is designed to predict the outcome of a single match based on various factors, such as:

- **Team Ratings**: Derived from historical match data, these ratings indicate the relative strength of the home and away teams.
- **Dominance Factor**: A parameter that measures the home team's advantage or disadvantage over the away team.
- **Goal Probabilities**: These are computed using the Poisson distribution based on the ratings and dominance factor.

```
def calculate_single_match( home_team , away_team ,
                            ratings_dict , implied_tg):

    home_rating , away_rating = get_ratings( home_team ,
                                             away_team ,
                                             ratings_dict)

    dominance = minimize_dominance( implied_tg ,
                                    home_normalized_prob)

    lambda_home = (dominance + implied_tg) / 2
    lambda_away = implied_tg - lambda_home

    probabilities = lmbda_to_probabilities( lambda_home ,
                                            lambda_away ,
                                            inflation = True
   )
    return probabilities
```

Listing 3.4: calculate_single_match function

Ratings Lookup: The function begins by retrieving the ratings for the home and away teams using the `get_ratings` function presented before. This provides a baseline for each team's strength, which will influence the goal expectations for the match.

Dominance Calculation: Next, the `minimize_dominance` function is called to compute the dominance factor for the home team. The dominance factor is used to adjust the expected goal-scoring rates, depending on whether the home team is stronger or weaker than the away team.

Poisson Goal Rates: Using the dominance factor and the implied total goals (TG), the function computes the expected goals (λ values) for both teams:

$$\lambda_{\text{home}} = \frac{\text{dominance} + \text{implied total goals}}{2}$$

$$\lambda_{\text{away}} = \text{implied total goals} - \lambda_{\text{home}}$$

Finally, the `lmbda_to_probabilities` function calculates the probability of each outcome (home win, draw, and away win) based on the goal rates from the Poisson model. The function considers various scorelines and computes the aggregate probabilities for each result.

3.2.5 Example Computation

To demonstrate how the model computes match probabilities and odds, let's consider the Manchester United vs. Chelsea match on 25/05/2023.

Match Data:

- Home Closing Odds: 1.55
- Draw Closing Odds: 4.39
- Away Closing Odds: 6.53

The `minimize_dominance` function computes the dominance factor for Manchester United, representing their advantage or disadvantage over Chelsea:

- **Dominance Factor**: 1.1271

The `get_ratings` function retrieves the ratings for both teams and the Home factor parameter:

- Home Team Rating: 0.5432
- Away Team Rating: 0.0736
- Home Factor: 0.3243

Using the dominance factor and the **implied total goals** (TG), the model computes the expected goals (λ) for both teams:

$$\lambda_{\text{home}} = \frac{1.1271 + 2.5}{2} = 1.8136$$

$$\lambda_{\text{away}} = 2.5 - 1.8136 = 0.6864$$

These λ values represent the expected goals for Manchester United and Chelsea, respectively.

The `lmbda_to_probabilities` function is then used to compute the probabilities for the match outcomes based on the Poisson-distributed goal rates:

- **Home Win Probability**: Based on the Poisson distribution for λ_{home} and λ_{away}, the probability of Manchester United winning is calculated.
- **Draw Probability**: Calculated as the probability that both teams score an equal number of goals.
- **Away Win Probability**: Based on the Poisson distribution for Chelsea's λ_{away}.

These probabilities are converted into betting odds:

- Home Win Odds: 1.79
- Draw Odds: 4.07
- Away Win Odds: 5.13

These calculated odds are close to the actual closing odds provided by bookmakers for this match.
This example provides a step-by-step breakdown of how the model computes the match probabilities and odds using the ratings and Poisson distribution.

Chapter 4

Model Evaluations

In this final chapter, the performance of the predictive model is going to be evaluated and discussed by applying well-known evaluation metrics from the literature. The goal is to assess the accuracy and reliability of the model in predicting match outcomes.

Following this, two distinct betting agents are introduced. Each agent will utilize the model's predictions to make informed betting decisions, staking on matches where the model identifies value. These agents are designed to simulate different betting strategies, and their performance will be analyzed in terms of profitability and loss, based on various parameters.

Finally, potential improvements to the model will be discussed, drawing insights from trials conducted throughout the research process. In this section possible refinements will explored and how they might enhance the model's performance. The chapter ends with a reflection on the results and suggestions for future work.

4.1 Evaluations

In this section, the evaluation of the predictive model will focus on the application of the Ranked Probability Score (RPS) [15] to assess its performance. As previously discussed in Chapter 2, various evaluation metrics were applied, such as the Brier Score and Ignorance Score, across different leagues and odds providers.

However, for the purposes of this section, we will concentrate solely on the RPS due to its unique properties that make it highly suitable for evaluating probabilistic predictions in football. The RPS is chosen here because it is both non-local and distance-sensitive, making it more appropriate for capturing the nuances in football

match predictions. Unlike local metrics that only compare the predicted probability of the actual outcome with the observed outcome, the RPS considers the entire probability distribution across all possible outcomes.
Moreover, its distance-sensitive nature ensures that predictions that were closer to the actual result are penalized less, making it a more refined evaluation tool.

The evaluation in this chapter is carried out using the same set of leagues and bookmakers previously used in Chapter 2, ensuring consistency in the analysis. The leagues included are those where the model was tested extensively, and the results will be presented in a comprehensive table that shows the RPS scores for each league, with a particular focus on the closing odds from the same bookmakers evaluated earlier.

	RPS			
League	**Pinna Op.**	**Pinna Cl.**	**Bet365 Op.**	**Model**
England Premier-League	0.198	**0.195**	0.198	0.197
France Ligue-1	0.204	**0.200**	0.205	0.206
Spain LaLiga	0.197	**0.194**	0.198	0.198
Italy Serie-A	0.194	**0.190**	0.195	0.193
Italy Serie-C-Group-A	0.220	**0.218**	0.221	0.222
Italy Serie-C-Group-B	0.200	**0.198**	0.203	0.203
Italy Serie-C-Group-C	0.207	**0.205**	0.208	0.211

Table 4.1: RPS of the Model for different leagues

The first significant observation from the evaluation results is that the best RPS scores are consistently obtained with Pinnacle's closing odds. This result agrees with the Efficient Market Hypothesis by Eugene Fama [12]. As discussed in chapter 2, by the time the match is about to begin, all available information has already been incorporated into the odds. This includes team lineups, injuries, weather conditions, and other relevant factors. Therefore, the closing odds provide the most accurate reflection of the match true probabilities.

In contrast, the model's predictions are made long before the match begins. Although the model can calculate probabilities based on historical odds data, it is not designed to account for real-time factors like sudden injuries, last-minute lineup changes, or shifts in betting sentiment. These are exactly the kinds of factors that Fama's EMH accounts for, demonstrating why closing odds outperform predictions made earlier. The model is based on patterns from long-term historical odds, so, it is limited to incorporate new information in the way that closing odds do.

It's interesting how, even if the model underperforms compared to Pinnacle's closing odds, it is competitive with, and in some cases even outperforms, Pinnacle's opening odds and other bookmakers like Bet365.
A crucial point emerges: despite not being able to predict real-time factors, the model is still able to perform such as the industry's leading bookmakers are able to do at an early stage. This suggests that the model has the potential to predict outcomes with a similar accuracy to the internal models used by bookmakers and, in certain scenarios, might even have an edge over them.

To explore whether the model can, in certain cases, take advantage of these opportunities, the next section will introduce two agents designed to exploit the situations where the model finds value. These agents will assess whether the model can generate a profit by betting on matches where the odds differ significantly from the model predictions.

4.2 Agent-Based Simulation

To further evaluate the model's ability to predict match outcomes and identify profitable betting opportunities, we introduce two agents that are designed to simulate different betting strategies. These agents not only utilize the model's predictions but are also capable of dynamically calculating odds for upcoming matches and making informed decisions based on those odds in real time.
Both agents leverage the model's ability to process historical data and compare the odds generated by the model with those offered by bookmakers.

Their main task is centered around identifying value in the market by determining when the odds provided by the bookmakers deviate significantly (over a treshold) from the model's calculated probabilities. By analyzing these discrepancies, the agents are able to compute the expected value (EV) of potential bets and decide whether or not placing a bet would be profitable.

4.2.1 Agents definition

The first agent operates in a straightforward manner. Each day, this agent calculates the odds for upcoming events using the model and compares them to those offered by the market. The agent makes betting decisions based on the Expected Value (EV), which is computed by comparing the model's odds with the bookmaker's odds.
The key factor guiding the agent's decision-making process is a predefined threshold. If the agent finds that the EV exceeds this threshold, it will place a bet using a flat staking system. The flat stake ensures that the agent bets a consistent amount on each match where the model identifies value, regardless of how large the perceived edge is. This approach simulates a simple betting strategy that focuses on finding situations where the potential reward justifies the risk, as determined by the threshold.

The second agent, takes a more sophisticated approach. Like the first agent, it uses the model to predict the odds for upcoming events and calculates EV. However, instead of using a flat staking system, this agent employs the Kelly Criterion to adjust the size of its bets based on the perceived edge.
The Kelly Criterion [16] is a well-known betting formula that helps determine the optimal stake size based on the edge and variance of the bet. By using this method, the agent can dynamically adjust its stake size depending on how confident it is in the value identified by the model.

The Kelly criterion formula is given by:

$$f^* = p - \frac{(1-p)}{b}$$

Where:

- f^* is the optimal fraction of the bankroll to wager,
- p is the probability of winning,
- $q = 1 - p$ is the probability of losing,
- b is the ratio of the amount won to the amount bet.

Alternatively, a more detailed version of the formula is:

$$f^* = p - \frac{(1-p)}{b}$$

This makes the second agent more adaptable and potentially more profitable in the long run, as it can exploit larger edges more aggressively while minimizing risk on smaller edges.

4.2.2 Agents evaluation

In this section, we analyze the performance of the agent_raw and agent_kelly—over a period of six years across top European football leagues. These leagues include Italy Serie A, England Premier League, France Ligue 1, Spain La Liga, and Italy Serie C. The agents operated under different thresholds (t1.15, t1.3, t1.5) to determine when it was profitable to place a bet, based on the calculated Expected Value (EV).

Each agent made decisions by comparing the model's calculated odds to the market odds. The first agent used a flat staking system, while the second agent, more sophisticated, applied the Kelly Criterion to adjust bet sizes dynamically based on the strength of the expected value.

The performance of each agent was measured in terms of several metrics:

- **Number of bets placed vs. available opportunities**: Indicating how selective the agent was based on the threshold.
- **Average EV**: Highlighting the value the model identified in the bets.
- **Average market odds vs. modeled odds**: Comparing the bookmakers' opening odds to the odds computed by the model.
- **Net profit and yield**: Showing the overall financial performance of the agent, expressed as both the total profit and yield (the percentage of returns on stakes).

The agents operated with additional constraints in the Italy Serie C league, where betting odds were limited to a maximum of 4 or 5, in order to reduce high variances in profitability. The table below summarizes the performance of the agents across leagues and thresholds, providing insight into how well each strategy performed under different conditions.

	Stats	**Threshold**		
		t = 1.15	**t = 1.3**	**t = 1.5**
Italy Serie A	Bets	334 / 1919	234 / 1919	171 / 1919
	Avg. EV	0.31	0.53	0.67
	Avg. betting odds	5.52	6.57	7.22
	Avg. model odds	3.91	4.13	4.88
Raw Agent	Bankroll	94.82	101.32	106.32
	Yield	-2.3%	2.1%	14.0%
Kelly Agent	Bankroll	98.64	100.76	104.97
	Yield	-1.1%	1.0%	6.6%

Table 4.2: Italy Serie A - Agent Performance Comparison with Yields

	Stats	**Threshold**		
		t = 1.15	**t = 1.3**	**t = 1.5**
England Premier League	Bets	288 / 1920	169 / 1920	51 / 1920
	Avg. EV	0.22	0.41	0.55
	Avg. betting odds	4.91	5.45	6.50
	Avg. model odds	3.39	4.56	5.51
Raw Agent	Bankroll	89.97	95.05	95.80
	Yield	-5.3%	-5.1%	-4.8%
Kelly Agent	Bankroll	92.55	95.98	97.39
	Yield	-4.3%	-4.0%	-3.1%

Table 4.3: England Premier League - Agent Performance Comparison with Yields

	Stats	**Threshold**		
		t = 1.15	**t = 1.3**	**t = 1.5**
Spain LaLiga	Bets	231 / 1107	167 / 1920	86 / 1920
	Avg. EV	0.41	0.60	0.71
	Avg. betting odds	4.82	6.29	6.98
	Avg. model odds	3.89	4.20	4.29
Raw Agent	Bankroll	99.92	103.5	105.8
	Yield	-0.0%	5.0%	13.0%
Kelly Agent	Bankroll	101.67	106.83	115.02
	Yield	0.1%	6.2%	9.4%

Table 4.4: Spain LaLiga - Agent Performance Comparison with Yields

	Stats	Threshold		
		t = 1.15	**t = 1.3**	**t = 1.5**
Italy Serie C	Bets	406 / 3121	311 / 3121	200 / 3121
	Avg. EV	0.27	0.36	0.45
	Avg. betting odds	4.00	4.56	5.71
	Avg. model odds	2.76	3.53	4.37
Raw Agent	Bankroll	186.83	211.63	246.01
	Yield	11.1%	17.6%	24.7%
Kelly Agent	Bankroll	199.54	254.10	290.12
	Yield	13.2%	19.6%	24.7%

Table 4.5: Italy Serie C - Agent Performance Comparison with Yields

	Stats	Threshold		
		t = 1.25 M = 4	**t = 1.25 M = 5**	**t = 1.3 M = 4**
Italy Serie C	Bets	32 / 3121	48 / 3121	17 / 3121
	Avg. EV	0.34	0.36	0.39
	Avg. betting odds	3.14	3.49	3.26
	Avg. model odds	2.35	2.57	2.36
Raw Agent	Bankroll	201.77	266.41	267.82
	Yield	15.7%	21.6%	35.0%
Kelly Agent	Bankroll	197.86	268.13	300.5
	Yield	14.0%	19.5%	24.9%

Table 4.6: Italy Serie C - Agent Performance Comparison with additional threshold

The results presented are grouped by league to offer a clearer comparison of how the agents perform in different contexts. The threshold (t) is applied to filter bets based on the Estimated Value (EV), which is calculated as the ratio of available odds to modelled odds. A threshold of 1.15, for example, means that only bets with at least a 15% higher EV than the market are considered.

One of the first observations is the significant difference between higher-profile leagues, such as Serie A, Premier League, and La Liga, and lower-tier leagues like Serie C. The agents seem to struggle considerably in the more competitive, higher-information leagues, whereas they perform far better in the less-publicized minor leagues.

The Premier League results highlight one of the most significant findings: despite finding good average EV (e.g., 0.22 with t = 1.15), the raw agent still experiences a decline in bankroll after 288 bets, resulting in a negative yield. This suggests that betting on this league, based on the model, may not be profitable in the long run. The reasoning behind this is clear—primary leagues like the Premier League are heavily scrutinized, and odds quickly adjust based on up-to-date information that the model is unable to account for in real-time. Therefore, the timing of the model's predictions, far before match time, hinders its ability to capitalize on market inefficiencies.
On the other hand, we can see that the kelly agent—despite also struggling in the Premier League—manages to mitigate losses better than the raw agent by dynamically adjusting the stakes based on the model's confidence in the predicted EV. This highlights the power of the Kelly Criterion in risk management, even in markets where the model struggles.

In Serie A, we observe similar struggles for both agents, especially at a low threshold of t = 1.15, where both agents show negative yields. However, as the threshold increases, the agents begin to perform better, with the raw agent turning a profit at a threshold of t = 1.5. This suggests that while the model may not be effective at identifying small-value opportunities (low EV), it becomes more reliable when the EV is significantly higher, allowing it to exploit potential bookmaker pricing errors at the opening of the odds market.
Interestingly, even though both agents start to show a profit at higher thresholds, the kelly agent remains more conservative, reflecting its ability to limit losses and still make gains.

In La Liga, the results are even more promising for the kelly agent. While the raw agent nearly breaks even at t = 1.15 (ending with a bankroll of 99.92), the kelly agent shows a small profit, with a yield of 0.1%. As the threshold increases to 1.3

and 1.5, both agents start to perform better, with the Kelly agent consistently outperforming the raw agent, indicating the Kelly agent's ability to capitalize on high EV situations while minimizing losses. The positive yields at higher thresholds suggest that the model performs better in La Liga when it is highly selective in its bets.

The most dramatic difference in performance appears in the lower-tier Serie C league. Both agents manage to outperform their results in the higher leagues, with notable bankroll increases and yields of 11.1% to 24.7% for the raw agent and 13.2% to 24.7% for the Kelly agent across the thresholds.
This can be attributed to the fact that minor leagues like Serie C are less scrutinized, with fewer professional models and bettors influencing the market. As a result, the model can better identify inefficiencies in bookmaker pricing, making this league a fertile ground for the application of the model.
The application of a maximum odds threshold (limiting the odds to 4 or 5) further enhances the agents' performance by reducing high-variance bets that could lead to unpredictable results. This additional constraint prevents the agents from betting on extremely high odds, which could be distorted due to external factors like end-of-season dynamics, where a team might have lost motivation or other unpredictable circumstances.

General Observations

Across all leagues, the threshold plays a critical role in filtering out low-value opportunities. As the threshold increases from 1.15 to 1.5, the number of bets decreases, but the average EV increases, leading to better overall performance and higher yields. The threshold essentially acts as a quality filter, ensuring that only the most promising opportunities are considered. This is particularly evident in Serie A and La Liga, where low thresholds result in small profits or losses, but higher thresholds lead to significant improvements in performance.

The Kelly agent consistently outperforms the raw agent in limiting losses and generating higher profits. This is particularly evident in leagues like La Liga and Serie C, where the Kelly agent not only avoids significant losses but also takes advantage of high EV situations, resulting in positive yields across most thresholds.

4.2.3 Ethical Considerations

The gross volume of gambling in Italy in 2023 increased by 10.2% compared to 2022, reaching 150 billion euros and setting a new record after the 136 billion recorded the previous year. The total value of bets exceeds 7% of the national GDP. It's worth noting that by 2021, the total collection had already returned to pre-pandemic levels, in a scenario where the first part of the year still saw restrictions aimed at containing the spread of COVID-19.
In the 2022-23 period, the recovery in the volume of physical gambling was evident (although it has not yet returned to 2019 levels), along with the continuous and significant expansion of online gambling platforms.
The per capita collection in 2023 for both physical and remote gambling — calculated based on the adult resident population in Italy, as recorded by ISTAT — was 2,996 euros (compared to 2,731.68 euros in 2022 and 2,229 euros in 2021).

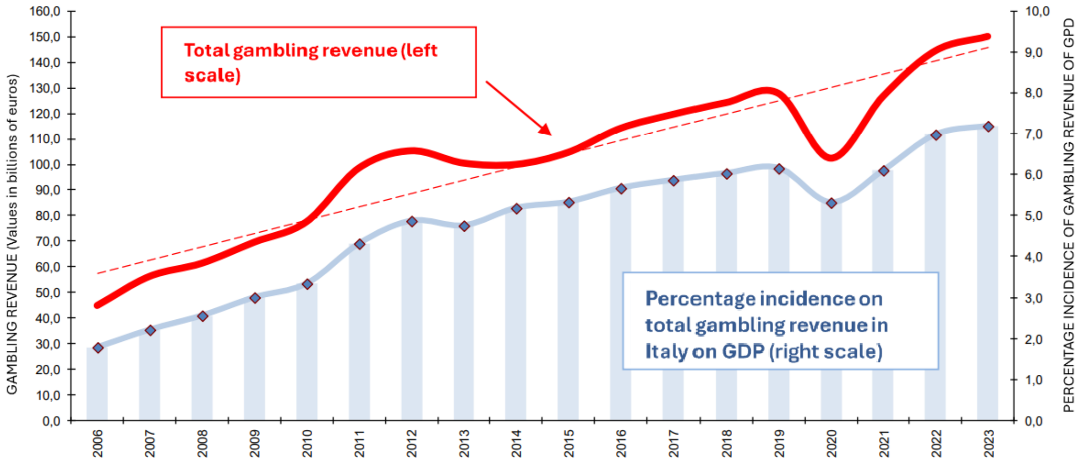

Figure 4.1: Percentage incidence of total gambling revenue in Italy on GDP and total gambling revenue – Absolute values in millions of euros (Real values – year 2023). Period 2006-2023.

The total amount wagered through physical networks was 67.9 billion euros (+7.8% compared to 2022). Online gambling reached 82.08 billion euros (+12.3% compared to 2022, particularly in card games, fixed-odds games, and sports-based games).

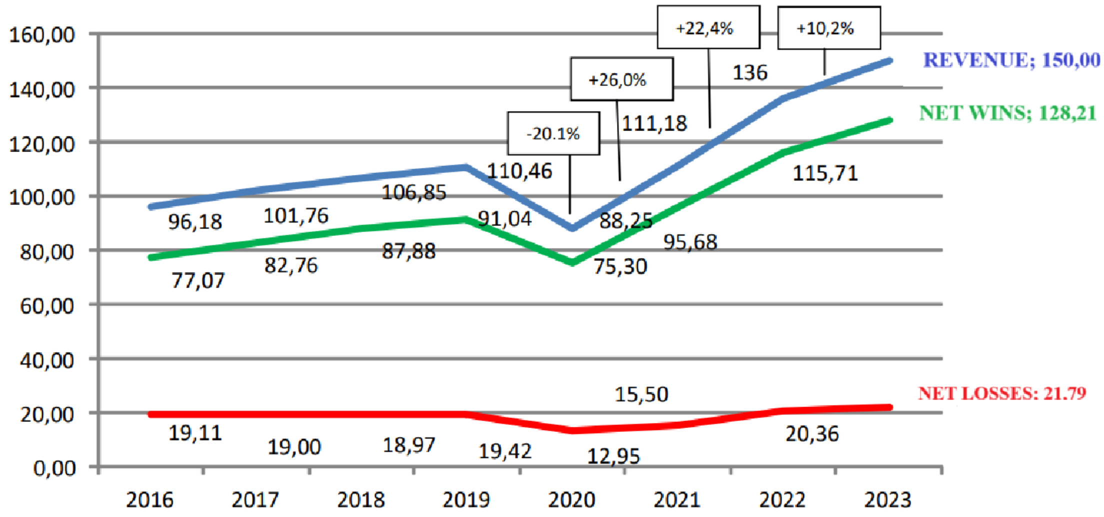

Figure 4.2: Revenue, winnings, and expenses (losses) recorded for the overall gambling sector. National data, Period 2016-2023 (*). Absolute values (in billions of euros) and percentage change compared to the previous year.

To truly understand the magnitude of 150 billion euros in gambling revenue, one must grasp the enormity of this figure. It amounts to 89% of the total food expenditure of Italians, as estimated for 2023. Additionally, 150 billion euros is five times the value of the 2024 national budget law. For context, healthcare spending in 2023 amounted to 131.1 billion euros.

As for the 22 billion euros lost by Italians in gambling, it is equivalent to completely wiping out the net annual income of over 1.1 million full-time workers with good wages and seniority, each earning a net monthly salary of around 1,500 euros. The figures would escalate even further when considering the average wages in certain sectors.

One of the main reasons I chose this thesis topic is because I am personally very concerned about this issue, and the data is self-explanatory. Italy holds several negative records in Europe: we rank near the bottom for internet connectivity, Wi-Fi availability, and network coverage, with less than 50% high-speed internet coverage. Yet, despite these technological shortcomings, Italy ranks among the top countries in the world for online gambling, particularly in the regions where internet access is weakest.

There is an inverse relationship between the country's financial socioeconomic situation and the increase in overall gambling. As the economic crisis intensifies—whether the crisis is real or perceived—gambling behavior increases, while consumer spending decreases. This dynamic, fueled by the widespread promotion of legal gambling, is driven by the illusory belief that a single win could solve financial problems in one stroke, especially in times of economic crisis.

The issue of gambling among young people is particularly concerning. A recent survey by Federconsumatori Modena [17], conducted among over a thousand high school students, revealed that one-third of them—most of whom were minors—are familiar with gambling. On average, three students per class (about 12% of the sample) view gambling as central to their future. There is no significant gender difference in gambling habits, although boys and girls are drawn to different types of games. It's becoming increasingly clear that certain video games, which are considered harmless, may actually prime young people for gambling from a young age. These games often simulate winning or require in-app purchases to continue playing, fostering an early fascination with gambling.
For younger generations, the shift from physical gambling to online gambling happened some time ago, and this makes them a crucial demographic for studying the gambling phenomenon and its implications. These young people gamble in their classrooms, in their bedrooms, and without the social control that exists in physical gambling environments. Prepaid cards allow them to gamble increasingly larger amounts, often unnoticed by their parents. Another growing concern is the allure of online trading, which in many ways mirrors gambling.
The rise of Twitch streams during the COVID-19 pandemic, where streamers would broadcast gambling sessions lasting two to three hours, further fueled the illusion of easy winnings among young viewers. While these streams were banned after two years, videos showcasing "highlight" wins remain on YouTube and other platforms, perpetuating the false impression that gambling leads to victory, when in reality, the only certainty in a casino is loss.

Compounding this issue is the inadequate attention from Italian legislation. While gambling ads have been banned, loopholes allow major gambling brands to continue appearing on television, often using domain extensions like .news or .live, specially created to evade restrictions.
Throughout my time at university and among my friends, I have witnessed firsthand how prevalent gambling is in many people's lives. From those who buy scratch-off tickets along with their cigarettes to those who hope to win the lottery after seeing someone win on TV, and even to those who place bets to "enhance their enjoyment" of a game—gambling is everywhere.
In my opinion, the root cause of this issue is: Ignorance. If you asked anyone, "Would you give me €100, and I'll give you back €90?" no one would agree. Yet, this is exactly what happens from a purely analytical perspective when considering a large enough sample of gambling data. Data beats emotions, every time.

4.3 Conclusions

This thesis began with an extensive review of existing models in the literature for forecasting outcomes in low-scoring games, specifically focusing on the modeling of football scores. By analyzing the contributions of foundational models such as those by Maher (1982) [7], Dixon and Coles (1997) [8], and ELO-based rating systems, a clear understanding of the strengths and limitations of these approaches was established. These models provided valuable insights, but they lacked some critical elements that have become essential in modern predictive modeling, particularly in utilizing external market signals.

The core innovation of this thesis lies in the introduction of historical market prices as a key feature of the prediction model. Drawing inspiration from Fama's Efficient Market Hypothesis [12], this approach leverages betting odds, which reflect a vast array of public and private information, to enhance the predictive accuracy. Unlike traditional models that primarily focused on in-game statistics or team ratings, this method uses the collective wisdom of market participants to predict future outcomes. The use of betting markets in this way is an entirely novel approach, not previously found in the literature.

Once the theoretical basis for using market data was established, an in-depth analysis was conducted to compare different market providers. Differences between bookmaker odds were examined, and a justification for selecting specific markets over others was provided, focusing on those that consistently offered more reliable and informative odds.

The model was then developed in a step-by-step manner, explaining each methodological choice, from the use of the Poisson distribution to the incorporation of time-decayed ratings and dominance measures. These methods allowed for a robust framework capable of accurately forecasting match outcomes by integrating data from historical betting odds with statistical modeling.
To further evaluate the model, two agents were implemented to simulate decision-making processes based on historical market conditions. These agents, through backtesting, demonstrated the model's capability to identify valuable opportunities in the betting markets, effectively exploiting situations where the market odds diverged from the model's predictions.

The results of the agents' operations were analyzed in depth, providing evidence that the model can successfully guide decision-making in real-world scenarios. The agents' performances confirmed that the model offers a strategic advantage, as they were able to capitalize on market inefficiencies.

Future Works

This thesis has laid the groundwork for several potential future research directions. Firstly, while soccer has been the focal point of this study, the model's framework is versatile and can be adapted to other low-scoring sports, such as hockey or baseball, or even games with different scoring dynamics. The principles outlined here, especially the reliance on historical market prices, are not restricted to soccer alone and could be applied in a variety of domains.
Another important avenue for future work involves the expansion and optimization of the dataset. While the current dataset provides a solid foundation for prediction, integrating a broader range of data sources, including live market prices or advanced game statistics, could further refine the model's accuracy.
Additionally, certain parts of the code may benefit from further optimization, allowing the model to process larger datasets more efficiently.

In conclusion, this thesis has successfully introduced and tested an innovative prediction model based on historical market prices, with promising results in the context of soccer. The integration of forecasting markets into predictive models opens up new possibilities for future research and practical applications in sports analytics and beyond.

Bibliography

[1] Fotios Petropoulos, Daniele Apiletti, Vassilios Assimakopoulos, et al. «Forecasting: Theory and Practice». In: *International Journal of Forecasting* 38 (2022), pp. 705–871. DOI: 10.1016/j.ijforecast.2021.11.001. URL: https://doi.org/10.1016/j.ijforecast.2021.11.001 (cit. on p. 1).

[2] Sajedeh Norozpour and Mehdi Safaei. «An Overview on Game Theory and Its Application». In: *IOP Conference Series: Materials Science and Engineering*. Vol. 993. Available under CC BY 3.0 License. IOP Publishing, 2020, p. 012114. DOI: 10.1088/1757-899X/993/1/012114 (cit. on p. 1).

[3] Jing Zhao, Fan Zhang, Chao Zhao, and Gang Wu. «The Properties and Application of Poisson Distribution». In: *Journal of Physics: Conference Series*. Vol. 1550. Available under CC BY 3.0 License. IOP Publishing, 2020, p. 032109. DOI: 10.1088/1742-6596/1550/3/032109 (cit. on pp. 2, 25).

[4] D. J. Morgan. «Review of Facts from Figures by M. J. Moroney». In: *The Economic Journal* 61.244 (1951). A review of the book Facts from Figures by M. J. Moroney, highlighting both its strengths and shortcomings in explaining statistical applications., p. 632 (cit. on p. 3).

[5] G. J. S. Ross and D. A. Preece. «The Negative Binomial Distribution». In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 34.3 (1985), pp. 323–335. DOI: 10.2307/2987659. URL: https://www.jstor.org/stable/2987659 (cit. on p. 3).

[6] I. D. Hill. «Association Football and Statistical Inference». In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 23.2 (1974), pp. 203–208. DOI: 10.2307/2347001. URL: https://www.jstor.org/stable/2347001 (cit. on p. 3).

[7] M. J. Maher. «Modelling Association Football Scores». In: *Statistica Neerlandica* 36.3 (1982), pp. 109–118. DOI: 10.1111/j.1467-9574.1982.tb00783.x (cit. on pp. 4, 5, 54).

[8] Mark J. Dixon and Stuart G. Coles. «Modelling Association Football Scores and Inefficiencies in the Football Betting Market». In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46.2 (1997), pp. 265–280. DOI: 10.1111/1467-9876.00065 (cit. on pp. 5, 30, 35, 54).

[9] Football-Data. *Rating Systems for Fixed Odds Football Match Prediction*. PDF document. Available from: https://www.football-data.co.uk. 2003. URL: https://www.football-data.co.uk (cit. on p. 8).

[10] Lars Magnus Hvattum and Halvard Arntzen. «Using ELO Ratings for Match Result Prediction in Association Football». In: *International Journal of Forecasting* 26.3 (2010), pp. 460–470. DOI: 10.1016/j.ijforecast.2009.10.002. URL: https://EconPapers.repec.org/RePEc:eee:intfor:v:26:y:2010:i:3:p:460-470 (cit. on p. 9).

[11] Fabian Wunderlich and Daniel Memmert. «The Betting Odds Rating System: Using Soccer Forecasts to Forecast Soccer». In: *PLOS ONE* 13.6 (2018), e0198668. DOI: 10.1371/journal.pone.0198668. URL: https://doi.org/10.1371/journal.pone.0198668 (cit. on p. 13).

[12] Colin Read. *The Efficient Market Hypothesists: Bachelier, Samuelson, Fama, Ross, Tobin and Shiller*. Palgrave Macmillan, 2012, p. 94 (cit. on pp. 13, 30, 43, 54).

[13] Oliver Hinz and Martin Spann. «The Impact of Information Diffusion on Bidding Behavior in Secret Reserve Price Auctions». In: *Information Systems Research* 19.3 (2008), pp. 351–368. DOI: 10.1287/isre.1080.0190. URL: https://doi.org/10.1287/isre.1080.0190 (cit. on p. 13).

[14] Erik Štrumbelj. «On Determining Probability Forecasts from Betting Odds». In: *International Journal of Forecasting* (2013) (cit. on p. 15).

[15] Edward Wheatcroft. «Evaluating Probabilistic Forecasts of Football Matches: The Case Against the Ranked Probability Score». In: *arXiv preprint arXiv:1908.08980* (2019). URL: https://arxiv.org/abs/1908.08980 (cit. on pp. 17, 42).

[16] Xinyi Kong. «Kelly Criterion and its Application». In: *Highlights in Business, Economics and Management EDMI 2023, Volume 9*. Capital Normal University, 2023, pp. 630–635 (cit. on p. 45).

[17] Federconsumatori and CGIL. *Il libro nero dell'azzardo: Mafie, Dipendenze, Giovani*. In collaboration with ISSCON Foundation. The report discusses the expansion of gambling in Italy, focusing on issues of organized crime, addiction, and young people. May 2024. URL: https://www.federconsumatori.it/wp-content/uploads/2024/05/report_azzardo_nazionale.pdf (cit. on p. 53).