Data-Driven Decision Making in Sports Betting: An Empirical Analysis of Machine Learning Approaches for the Outcome Prediction of English Premier League Football Matches

Bachelor Thesis

Bachelor of Science in Software Systems Science

Di Bao

March 31, 2026

Supervisor:

1st: Prof. Dr. Christian Ledig

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

The global sports betting market is substantial and growing rapidly. Artificial intelligence and analytics offer investors a new, data-driven perspective. This study aims to build a betting framework by analyzing machine learning methods. It integrates outcome-prediction models with profitable betting strategies. The dataset contains over 5,000 real-world English Premier League matches from the 2010-2025 seasons. From this data, 31 quantitative features were created, including team-strength metrics such as Elo ratings. Offensive and defensive indicators for expected goals are also included. Seven mainstream machine learning models were tested. Instead of accuracy, we used the Ranked Probability Score (RPS) to more precisely capture the difference between predicted probability distributions and actual outcomes.

The evaluation framework showed that Random Forest models reached the highest predictive performance. Their RPS values matched or exceeded bookmaker benchmarks. After validating predictive ability, we designed five sets of comparison experiments to test different real-world betting scenarios. Empirical results showed that an anti-home strategy—filtering out extremely high and low odds—gave the highest ROI. Finally, we introduced the Fractional Kelly Criterion for dynamic capital allocation in simulations. This method adjusts the stake for each match based on the model's predicted edge. Backtesting results confirm that, for high-value betting opportunities, the compound returns generated by the Kelly Criterion far exceeded those of traditional flat betting. To bridge the gap between theoretical modeling and practical implementation, this research employed a Season-by-Season Walk-Forward Validation framework to simulate a dynamic investment environment over a seven-year horizon (2018–2024). This approach enabled continuous model retraining to adapt to market regime shifts. This research thus offers a reliable theoretical basis and a practical path for data-driven sports betting.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
EPL	English Premier League
ANNs	Artificial Neural Networks
KNN	K-Nearest Neighbors
ODM	Offense-Defense Model
ROI	Return on Investment
MDD	Maximum Drawdown
RFE	Recursive Feature Elimination
PCA	Principal Component Analysis

Notation

Please list all mathematical notations that are used in the thesis here. For reference, we provide you with an example from the book by **?**.

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix

Sets and Graphs

$\mathbb{R}$	The set of real numbers

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a discrete variable
E	Mathematical expectation
$Pois(\lambda)$	Poisson distribution

Functions

$\sum X_i$	Summary

Datasets and Distributions

$\mathcal{D}$	A data set
$\mathcal{D}_{train}$	A training data set
$\mathcal{D}_{test}$	A testing data set
$\mathbb{X}$	A set of training examples
$\boldsymbol{x}^{(i)}$	The i-th example (input) from a dataset

1 Introduction

1.1 Background

The global sports betting market is undergoing unprecedented expansion. According to recent industry data, the global sports betting market was valued at USD 191.66 billion in 2025 and is expected to reach USD 385.8 billion by 2033. This data is from databridge-marketresearch.com (2025). Football is undoubtedly the most popular sport worldwide. The English Premier League (EPL) constitutes a central component of the global sports betting market. Due to high viewership and international visibility, EPL matches account for a significant share of total betting turnover. Historically, most bettors used subjective intuition, team reputation, or expert recommendations. Recently, data analytics techniques have fundamentally transformed this traditional approach. Mandadapu (2024) notes that football betting has evolved from notepad-based predictions in the 1960s to a financial derivatives market that increasingly relies on artificial intelligence and machine learning. Bookmakers invest heavily in developing predictive models to set odds. Professional investors use large-scale datasets to find mathematical edges that can outperform the market. Data-driven decision-making is now a core principle in modern sports betting.

1.2 Problem Statement

Within this large and active market, an objective reality persists: the vast majority of retail bettors ultimately incur financial losses. The main cause of this widespread unprofitability is not that their predictions of match outcomes are entirely inaccurate, but rather the margin embedded in the odds offered by bookmakers. The margin represents an implicit fee, deducted from the odds, to ensure the bookmaker's risk-free profitability.

A simple coin-toss example helps explain this. The true probability of a fair coin landing on heads is 50%. Under fair conditions, the corresponding odds would be 2.0. To add their margin, bookmakers usually offer odds of just 1.90. If a bettor always bets at 1.90 odds with perfect 50% accuracy, their principal will decrease over time because of this 0.1 difference.

Therefore, in sports betting prediction research, it is logically flawed to determine the outcome only on the basis of model accuracy. High prediction accuracy does not necessarily translate into high financial returns. The true key to profitability lies in identifying and exploiting value bets. A value bet occurs when the odds offered by a bookmaker exceed the fair odds implied by the actual probability of an event. For example, if a model estimates that Arsenal's true probability of winning is 60%, corresponding to fair odds of approximately 1.67, but a bookmaker offers odds of 2.00 (implying a probability of 50%) due to public bias, a value margin of 10% emerges. The central research question of this study is: how can machine learning models be leveraged to accurately identify mispriced market opportunities and translate them into actual investment returns?

1.3 Research Purposes and Contributions

The main goal of this study is to build a comprehensive betting framework. The framework covers the full process, including data processing, model training, betting strategy, and capital management. Using historical English Premier League data, the study examines machine learning profitability under real-world market conditions. The paper offers four core contributions.

Contribution 1: A Feature Engineering Process with a Dual-Track Feature Selection Mechanism. Sports match data are highly susceptible to a critical error: future data leakage. To address this, this study proposes a rolling data-processing procedure that strictly follows chronological order. A suite of feature engineering techniques captures team strength from multiple perspectives. The goal is to approximate the informational efficiency of a theoretical "perfect model." Furthermore, to accommodate the specific features of different machine learning algorithms, a dual-track feature dimensionality reduction approach is used. Recursive Feature Elimination (RFE) is employed to select core features for tree-based models. Principal Component Analysis (PCA) reduces multicollinearity for linear models.

Contribution 2: Multi-Model Evaluation and Rigorous Market Benchmarking. Seven leading machine learning models, such as XGBoost and Random Forest, are built and compared. For evaluation, this study avoids possibly misleading accuracy metrics. Instead, it uses the Ranked Probability Score (RPS) as a more robust measure. The RPS of the top model is then directly compared with that of the actual betting market after removing the bookmaker margin. This comparison tests whether the models process information as effectively as, or even better than, bookmaker pricing systems.

Contribution 3: An Empirical Comparison of Betting Strategy Profitability. Instead of wagering on all potentially profitable matches, this study analyzes three different betting strategies to find the best way to capture value bets. Next, it compares two decision-making logics. Then, five betting strategies are designed based on various real-world operations. By comparing the returns and maximum drawdowns of each, the analysis finds a "sweet spot" that balances risk and return.

Contribution 4: An Empirical Back-Testing with Dynamic Capital Management. In the final stage, this study uses the Fractional Kelly Criterion—a well-known concept in finance—to dynamically manage capital. The Kelly Criterion adjusts the amount staked on each wager according to the model's predicted edge. Detailed back-testing shows that, for high-quality betting opportunities, the Kelly Criterion amplifies long-term returns through compounding.

2 Related work

The theoretical foundation of this study spans multiple disciplines, including computer science, statistics, and financial mathematics. To establish the academic positioning of

this research, a comprehensive literature review is conducted. This review covers the application of machine learning in sports prediction, the evolution of feature engineering, and investment strategies within betting markets.

2.1 Machine Learning in Sports Prediction

Academic research on sports outcome prediction has undergone a profound transition from traditional statistical methods to modern machine learning approaches. In their comprehensive literature review, Bunker and Thabtah (2019) note that early studies predominantly relied on logistic regression or simple Markov models. With advances in computational power, researchers have increasingly shifted toward nonlinear predictive frameworks. Baboota and Kaur (2019) provides strong empirical support in their study of the English Premier League, demonstrating that ensemble algorithms, such as Random Forests, significantly outperform single baseline models in identifying the complex patterns associated with match outcomes.

Other powerful tree models are also widely used. These include eXtreme Gradient Boosting (XGBoost), which combines many simple models called decision trees and improves prediction accuracy through repeated corrections. Berrar et al. (2019) combined with two machine learning models, a k-NN and an extreme gradient boosted tree (XGBoost), to predict a match outcome. Categorical Boosting (CatBoost) is designed to work with categorical data—meaning fields that can take on one of a set of values, such as "home team name" or "referee name" in football data. Malamatinos et al. (2022-08-31) got the highest accuracy with CatBoost model in his study. CatBoost can use these kinds of data directly, eliminating the need to convert them into numbers first (a process known as manual encoding). To create a broad comparative baseline, additional models—including Artificial Neural Networks (ANNs) Bunker and Thabtah (2019), inspired by the way brains process information, and Naïve Bayes (NB), a probabilistic model based on Bayes' theorem, and K-nearest neighbors (KNN)Brooks et al. (2016)—are included. This approach enables an objective evaluation of the most suitable predictive engine for sports betting applications.

2.2 Data processing

Raw football match records, such as scores, red cards, and yellow cards, contain substantial noise. Consequently, feature engineering constitutes a critical step in refining model predictive accuracy. Among the various features employed, dynamic rating systems occupy a central role.

The Elo rating system was developed by Elo (1978) to evaluate chess player strength. Today, data scientists widely use it for football prediction. Robberechts and Davis (2019) used ELO ratings in predicting the result of FIFA World Cup matches. Its core logic is intuitive: team ratings adjust based on how actual match results compare with expected win probabilities, which reflect each team's likelihood of winning given its current ratings. For example, if two teams start with a baseline rating of 1500 points, their expected win probabilities are equal. When a lower-rated team (1200 points) beats a higher-rated team (1800 points), the underdog gains a substantial rating increase while the favorite suffers a significant penalty. In this study, the HomeElo and EloDiff features use this mechanism to capture short-term fluctuations in team form; specifically, HomeElo represents a team's

Elo rating prior to a match, and EloDiff is the difference in Elo ratings between the home and away teams.

Furthermore, Koopman and Lit (2019-04) highlights the importance of score-driven models, which update their estimates as new data arrive, for time-series forecasting. Statistical studies show that the number of goals scored per football match roughly follows a Poisson distribution, a probability model suitable for counting rare events. Based on this, an Offensive/Defensive Model (ODM) is constructed. This model calculates the ratio of a team's average goals scored to the league's average. An offensive ratio (HomeOff) above 1.0 signals greater attacking strength than the league average. The Dolores prediction model, developed by Stübinger et al. (2019-12-19) and Constantinou (2019-01), shows that using multidimensional historical performance data—such as recent goal differential (the difference between goals scored and conceded in recent matches) and home/away-specific win rates—greatly improves model stability and predictive breadth.

2.3 Betting Markets, Calibration, and Profitability

The efficient market hypothesis in finance states that market prices reflect all public information. In sports betting, the odds set by bookmakers serve as authoritative market prices for a match. Bookmakers have large data analytics teams and access to insider information. Their odds are generally considered highly accurate. Beating this benchmark is a tough challenge.

To evaluate if predictive models can outperform bookmakers, prior research often focused on high prediction accuracy. Walsh and Joshi (2024)question this focus and argue that, in real-money sports betting, a model's calibration is more important than accuracy. For instance, if a model predicts a 70% win probability, the team should win about 70 out of 100 similar cases. Berrar et al. (2024-10) further stresses that modern algorithms should output probability vectors for win, draw, and loss outcomes, rather than just binary labels.

Following this perspective, the present study does not use accuracy as the main evaluation metric. Instead, it adopts the Ranked Probability Score (RPS) as the core measure of model performance. The RPS sums the squared errors between the model's predicted cumulative distribution and the actual match result. Lower RPS values indicate better model performance. This metric penalizes "highly confident but incorrect" predictions, making it the industry standard for evaluating professional betting models.

Finally, value identification and capital allocation form the final closed loop of quantitative investment. Wilkens (2021) finds that wagering solely on favorites in tennis is unlikely to produce long-term profit. Stübinger and Knoll (2018) shows that achieving a positive ROI in football betting requires identifying value bets. In this study's strategy, value is defined as the probability difference. A betting signal occurs only when the model's estimated win probability is much higher than the bookmaker's implied probability.

Once a profitable set of betting opportunities is identified, the question of how to allocate stakes is critical. Kelly (1956) introduced the Kelly Criterion, a mathematical formula that determines the optimal fraction of capital to wager when both the probability of winning and the betting odds are known. Wagering according to this criterion theoretically maximizes the long-term compound growth of capital. However, due to the inherent uncertainty in sports outcomes, full Kelly wagering may result in significant drawdowns.

Therefore, a Fractional Kelly strategy within its back-testing framework is a better choice. Matej et al. (2021) shows that the practical necessity of the additional risk-control methods with the Kelly criterion and demonstrates their individual benefits. In particular, the implementation may wager only one-quarter (25%) of the amount suggested by the full Kelly formula. This conservative approach to dynamic capital management seeks to capture the benefits of compounding while establishing a robust risk-control mechanism for the investment portfolio.

Table 1 includes descriptive information, such as the main findings, data used, and models, from the related literature.

Table 1: Related work using machine learning techniques to predict sports matches

Author	Modeling techniques	Data	Main findings
Baboota and Kaur (2019)	Gaussian Naive Bayes, Support Vector Machine, Random Forest, Gradient Boost	English Premier League data from 2005 to 2016 and FIFA database for rating statistics	Gradient Boosting performs best with 56.7% accuracy and RPS of 0.2156
Stübinger and Knoll (2018)	Random Forest, Boosting, Support Vector Machine, Linear Regression	Data from the five top European leagues from season 2013 to 2017	Prediction accuracy of around 75.62% achieved with Random Forest
Walsh and Joshi (2024-06)	Linear Regression, Random Forest, Support Vector Machine, Multilayer Perceptron	Five NBA seasons from 2014/2015 to 2018/2019	The most accurate predictive model is Linear Regression, accuracy: 69.23%
Wilkens (2021)	Linear Regression, Random Forest, Neutral Network, Support Vector Machine, Gradient Boost	Approximately 39,000 professional tennis singles matches that were played between 2010 and 2019, across ATP and WTA tournaments	Several models were very close, but Random Forest slightly outperformed them with 70.9%
Stübinger et al. (2019-12-19)	Linear Regression, Random Forest, Boosting, Support Vector Machine, and an equal-weight ensemble ALL that integrates four models	The data comes from the official database of the football simulation game FIFA and includes more than 40 specific attributes of each player, spanning from the 2006/2007 season to the 2017/2018 season	The equally weighted ensemble model performed best, with a prediction accuracy of 81.77%, followed closely by the random forest model at 81.26%

(Continued on next page)

Table 1: *(Continued from previous page)*

Author	Modeling techniques	Data	Main findings
Mandadapu (2024)	Random Forest, XGBoost, Support Vector Machine, KNN	This dataset includes detailed stats from the English Premier League seasons of 2021-2022 and 2022-2023	When trained using all features and single-season data, SVM achieved a prediction accuracy of 72.67%
Constantinou (2019-01)	Dolores: a model with a mixture of two methods: dynamic ratings and Hybrid Bayesian Networks	The dataset is provided for the competition Machine Learning for Soccer LOPES et al. (2017) The data contains incorporates 216,743 match instances from different football leagues throughout the world	The Dolores model achieved an RPS of 0.208256
Berrar et al. (2024-10)	KNN, ANN, Ordinal Forest, Naive Bayes	Data provided by the 2023 Soccer Prediction Challenge, which comes from 51 leagues from 34 countries, from 2000 to 2023	ANN performed best, with its accuracy (RPS = 0.2113) being very close to Bookmakers' model
Ren and Susnjak (2022)	Logistic Regression, Random Forest, KNN, Gradient Boost, CatBoost	English Premier League data from season 2001 to season 2021	There is no single optimal model that works for all situations, Random Forests offer the best overall performance
Malamatinos et al. (2022-08-31)	KNN, Random Forest, Support Vector Machine, LogitBoost, CatBoost	Data from the Greek league covering six seasons from 2014 to 2020	CatBoost achieved the highest accuracy of 67.73%.

3 Methodology

This section gives a detailed account of the original data sources and the challenges met during data cleaning. It also explains the step-by-step process of feature engineering, the underlying mathematical principles, and the implementation of feature selection.

3.1 Data Collection and Time-Series Split

Historical match data for the English Premier League (EPL) were collected from publicly available football statistics databases Football-Data.co.uk (2025). The dataset spans from August 2010 to October 2025 and comprises 5,790 match records. Each record includes information such as match date, home team name, away team name, full-time score, and the original odds offered by multiple bookmakers.

A central challenge in processing these data is the risk of temporal causality inversion, which means mistakenly allowing information about future events to influence predictions about the past or present. In sports prediction research, a model that inadvertently accesses future match outcomes during training suffers from data leakage, the unintended flow of information from the future to the model's learning process. Such leakage leads to artificially inflated performance in laboratory settings but inevitably results in failure in real-world investment scenarios.

To eliminate this risk, a strict unidirectional temporal ordering strategy was implemented. All 5,790 records were sorted by match date. July 2024, during the summer off-season, served as the natural temporal dividing line. Matches from August 2010 to May 2024 formed the training set. Machine learning models learned patterns only from this historical period. Matches from August 2024 to October 2025 made up the test set, simulating an unknown future investment environment. This rigorous out-of-sample testing ensures all return-on-investment (ROI) conclusions are objective and replicable. Let $\mathcal{D}$ be the complete dataset ordered by date t. The split is defined as:

$$\mathcal{D}_{train} = \{d \in \mathcal{D} \mid t(d) < T_{split}\}, \quad \mathcal{D}_{test} = \{d \in \mathcal{D} \mid t(d) \geq T_{split}\} \tag{1}$$

where T_{split} corresponds to July 2024.

3.2 Feature Engineering Construction

Raw match scores and odds contain substantial background noise that machine learning models cannot directly interpret in tactical terms. The original features can be seen in Table 2. Feature engineering, therefore, constitutes the most critical step in enhancing model predictive capability. A total of 31 core quantitative features were constructed. When calculating a team's historical statistical features, a one-step lag was enforced in the code, ensuring that, when predicting the tenth match, the model could only access aggregated data from the team's previous nine matches.

These 31 features are categorized into four logical modules. Their names and descriptions are given in Table 3.

3.2.1 Market Implied Probability Features

The decimal odds offered by bookmakers contain dense market information. These odds were first converted into implied probabilities. For example, if the odds for a home win are 2.0, the raw implied probability is 50%. Bookmakers include a profit margin, known as overround, in the odds to ensure risk-free profitability. As a result, the sum of raw probabilities for the three outcomes—home win, draw, and away win—usually totals about 105

To extract the pure market consensus, a normalization formula was applied to remove this margin τ, where $\sum P_{raw} = 1 + \tau$, we apply normalization:

$$P_{true,i} = \frac{1/O_i}{\sum_{j \in \{H,D,A\}} (1/O_j)} \tag{2}$$

Table 2: Description of the original features

Feature name	**Description**
Date	Date of the game
HomeTeam	Name of the Home team
AwayTeam	Name of the away team
FTHG	Full time home goals
FTAG	Full time away goals
FTR	Full time result
HTHG	Half time home goals
HTAG	Half time away goals
HTR	Half time result
Referee	Name of the Referee
HS	Home shots
AS	Away shots
HST	Home shots on target
AST	Away shots on target
HF	Home fouls
AF	Away fouls
HC	Home corners
AC	Away corners
HY	Home yellows
AY	Away yellows
HR	Home reds
AR	Away reds
B365H	Bet365 home odds
B365D	Bet365 draw odds
B365A	Bet365 away odds

where $i \in \{H, D, A\}$ represents Home Win, Draw, and Away Win respectively. The resulting true probabilities were denoted as ImpPH, ImpPD, and ImpPA. As demonstrated by Walsh and Joshi (2024), these margin-adjusted probability features provide machine learning models with a highly robust benchmark anchor for prediction.

3.2.2 Dynamic Elo Rating System

The competitive strength of football teams fluctuates continuously. The Elo rating system, originally developed by Elo (1978) for evaluating chess players, was introduced to quantify such dynamics. Initial Elo ratings for all teams were uniformly set at 1,500 points. After each EPL match, the system compared the actual match result with the pre-match expected win probability to incrementally adjust team ratings.

The expected outcome E_A for Team A against Team B is given by the sigmoid-inverse function:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A + \gamma)/400}} \tag{3}$$

where $\gamma = 100$ represents the home-field advantage constant. The ratings are updated post-match:

$$R_{A,new} = R_{A,old} + K(S_A - E_A) \tag{4}$$

where $S_A \in \{1, 0.5, 0\}$ is the actual match result and $K = 20$ is the learning rate. This mechanism allows teams to gain or lose many rating points after unexpected outcomes. The features HomeElo and EloDiff come from this system and help the model assess the likelihood of upsets.

3.2.3 Offensive and Defensive Modeling (ODM)

The number of goals scored in a football match is not purely random. Time-series research by Koopman and Lit (2019-04) confirms that goal counts follow a Poisson distribution. This theoretical foundation led to the construction of an Offensive/Defensive Model (ODM).Assuming goal counts follow a Poisson distribution $X \sim \text{Pois}(\lambda)$, we define the relative strength ratios:

$$\text{Off}_i = \frac{\frac{1}{n}\sum_{k=1}^{n} G_{g,i,k}}{\bar{G}_{league}}, \quad \text{Def}_i = \frac{\frac{1}{n}\sum_{k=1}^{n} G_{c,i,k}}{\bar{G}_{league}} \tag{5}$$

where G_g and G_c are goals scored and conceded, respectively.

An offensive ratio (HomeOff) over 1.0 shows a team's attacking strength surpasses the league average. A defensive ratio (HomeDef) was also calculated using goals conceded. These features translate tactical matchups into specific numerical ratios.

3.2.4 Historical Form and Dynamic Rolling Features

Recent changes in team fitness and morale affect match outcomes. Rolling window time-series techniques were used. Dynamic form is captured using a sliding window of size $w = 6$:

$$\text{AvgDiff}_t = \frac{1}{w}\sum_{i=1}^{w}(X_{t-i} - Y_{t-i}) \tag{6}$$

where X and Y represent team-specific statistics (e.g., shots, corners). For each team, averages from their latest six matches were extracted. Indicators included features such as average

goal differential and average shots-on-target differential (e.g., AvgGDiff). Implementation, a grouping function (groupby('Team')) was strictly applied, ensuring that data for the 20 EPL teams were physically isolated during calculations. This isolation completely prevented systemic errors, such as incorrectly aggregating Liverpool's shot data into Manchester United's historical features.

3.2.5 Data Continuity and Truncation Strategy for Initial Matches

In time-series feature engineering, the treatment of the season's opening matches presents a significant methodological challenge. In many conventional sports analytics frameworks, cumulative statistics (such as league points and goal averages) are reset to zero at the beginning of each new season. However, such a reset mechanism creates a cold-start problem: the first match of each team in each season would contain null values for all rolling and expanding features, leading to a substantial loss of sample size if those records are discarded.

To preserve the dataset's statistical power while maintaining predictive integrity, this study implemented a Cross-Season Continuous Accumulation Strategy. This approach is based on two core rationales:

Temporal Stability of Team Strength: The off-season in the English Premier League typically lasts only two to three months. From a dynamical systems perspective, the latent competitive strength and tactical identity of a professional squad do not vanish during this brief hiatus. Therefore, utilizing the trailing data from the end of the previous season to predict the start of the next provides a more stable and accurate representation of a team's true baseline than an arbitrary reset to zero.

Minimization of Systematic Bias: If a reset strategy were applied across the 15-year span (2010–2025) for 20 teams, at least 600 high-value match records would be invalidated due to missing historical antecedents. By treating the entire 15-year period as a continuous timeline, we encounter the null feature issue only at the first appearance of a team in the global dataset (i.e., their first match since 2010 or their first promotion to the EPL).

Consequently, a Selective Truncation Method was applied: only the absolute first match of each team in the entire longitudinal dataset was removed. This ensures that every match included in the training and testing phases possesses a valid, non-null historical profile (e.g., AvgGoals, EloDiff). This strategy optimizes the trade-off between data volume and feature quality, ensuring that the machine learning models are trained on a high-density, mathematically complete feature matrix.

Table 3: Descriptions of the candidate features

Feature	**Description**
ELO-based Features	
HomeElo	Elo ratings for the home team before the match
AwayElo	Elo ratings for the away team before the match
EloDiff	Elo ratings difference between home team and away team
EloHomeW	Probability of home team winning converted by ELO ratings

(Continued on next page)

(Continued from previous page)

Feature	Description
EloAwayW	Probability of away team winning converted by ELO ratings
EloDraw	Probability of a draw occurring converted by ELO ratings
ODM-based Features	
HomeOff	Home team's offensive capabilities
AwayOff	Away team's offensive capabilities
HomeDef	Home team's defensive capabilities
AwayDef	Away team's defensive capabilities
League point-based Features	
HomeP	Current home team points
AwayP	Current away team points
PDiff	Current points difference between home team and aways team
Statistical Features	
HomeW	Home team's win rate in all previous matches
HomeD	Home team's draw rate in all previous matches
HomeHW	Home team's win rate in home matches
HomeHD	Home team's draw rate in home matches
AwayW	Away team's win rate in all previous matches
AwayD	Away team's draw rate in all previous matches
AwayAW	Away team's win rate in away matches
AwayAD	Away team's draw rate in away matches
Form-based Features	
AvgGDiff	Average goal difference between two teams in previous six matches
AvgSDiff	Average shot difference between two teams in previous six matches
AvgCDiff	Average corner difference between two teams in previous six matches
AvgSTDiff	Average on target shot difference between two teams in previous six matches
AvgFDiff	Average fouls difference between two teams in previous six matches
HomeRestDays	Days since the home team's last match
AwayRestDays	Days since the away team's last match
Odd-based Features	

(Continued on next page)

(Continued from previous page)

Feature	Description
ImpPH	Implied possibility of home team win offered by the bookmakers
ImpPA	Implied possibility of away team win offered by the bookmakers
ImpPD	Implied possibility of draw offered by the bookmakers

3.3 Predictive Algorithms and Evaluation

After feature construction, seven machine learning algorithms were chosen for modeling. Selection criteria included their wide use in classification, frequent mention in sports prediction literature, and varied modeling approaches. This comparison aims to find the best algorithm for predicting EPL match outcomes.

Logistic Regression Logistic regression is a linear model used for binary and multiclass classification. The algorithm maps the output of linear regression into probability space to enable classification. Its main advantage is strong interpretability, allowing clear identification of each feature's effect on predictions. Logistic regression is robust to noise and performs well on small datasets. In football prediction, it is often the baseline model for evaluating the added value of complex algorithms.Beal et al. (2021)

Naive Bayes The Naive Bayes algorithm is based on Bayes' theorem and assumes conditional independence among features. Despite frequent violations of this assumption in real-world data, the algorithm still performs well in many applications. Its advantages include computational capability, scalability to high-dimensional data, and suitability for small sample sizes. In sports prediction, it is an efficient baseline, especially when features are relatively independent.Zhang (2005)

Artificial Neural Network Artificial neural networks simulate the structure of biological neurons. They construct multi-layer nonlinear transformations to approximate complex functional relationships. The architecture has input, hidden, and output layers. Network weights are adjusted via backpropagation to minimize prediction error. Neural networks fit nonlinear patterns and capture deep feature interactions. However, they require much training data and complex model tuning, which can lead to overfitting.

k-Nearest Neighbor The k-nearest neighbor (KNN) algorithm is an instance-based, non-parametric method. It makes no data distribution assumptions. To classify a sample, it calculates its distance to all training instances and selects the majority class from its k closest neighbors. KNN is easy to implement, has no explicit training phase, and handles multiclass problems. However, its computational cost increases with the size of the training set, and it suffers from the curse of dimensionality.Esme and Kiran (2018)

Random Forests Random forests are ensemble learning algorithms that construct multiple decision trees and aggregate their votes for classification. For each tree, bootstrap sampling is used to draw random samples from the original data, and at each node split, a random subset of features is considered. This dual randomization mechanism effectively reduces overfitting and achieves stronger generalization performance than single decision trees. The algorithm also outputs feature importance scores, facilitating feature selection. In sports prediction research, random forests have been widely adopted for their ability to detect nonlinear relationships.Li and Mu (2024)

XGBoost XGBoost is an efficient implementation of gradient boosting that introduces several optimizations to the traditional framework. It employs parallelized tree construction, substantially improving training efficiency. Regularization terms are incorporated into the objective function to control model complexity and prevent overfitting. At each iteration, the algorithm fits the residuals from the previous iteration, progressively optimizing predictions. XGBoost performs well on structured tabular data and has demonstrated excellent predictive performance across numerous machine learning competitions and practical applications.Chen and Guestrin (2016)

CatBoost CatBoost is a gradient boosting algorithm specifically optimized for categorical variables. Unlike XGBoost, CatBoost can directly process raw categorical features—such as "home team name" or "referee name"—without requiring manual numerical encoding. The algorithm adopts symmetric tree models and ordered boosting mechanisms to effectively mitigate prediction shift, a common issue in gradient boosting methods. CatBoost demonstrates strong robustness against overfitting and is particularly well-suited for sports data containing rich categorical information.Prokhorenkova et al. (2018)

3.4 Feature Selection

Including too many features in a machine learning model is not necessarily beneficial. Redundant features may introduce irrelevant noise, leading to overfitting. To address feature redundancy and the curse of dimensionality, a dual-track feature selection approach was implemented, tailored to the underlying mechanisms of different algorithm types.

3.4.1 Recursive Feature Elimination (RFE)

Tree-based models, such as random forest and XGBoost, are adept at handling complex nonlinear relationships among features. For these nonlinear models, Recursive Feature Elimination (RFE) was employed. The RFE process operates iteratively like a competitive elimination process. First, a base random forest model was constructed and trained on the full set of 31 features, producing importance scores I. RFE ranks features, and then the set of features F is pruned iteratively:

$$F_{next} = F_{current} \setminus \{f \in F_{current} \mid I(f) = \min(I)\} \tag{7}$$

until $|F| = 15$.

Ultimately, RFE retained the 15 most valuable features for prediction. The resulting feature importance rankings revealed that market implied probability (ImpPH) and Elo rating differential (EloDiff) dominated prediction outcomes, consistent with the economic principles that betting markets are highly efficient and that team strength determines fundamental performance.

3.4.2 Principal Component Analysis (PCA)

Linear classification models, such as logistic regression, impose strict requirements on input data. When a dataset contains highly correlated features—such as "home historical win rate" and "team overall win rate"—linear models have difficulty assigning stable weights. This issue is known as multicollinearity Kabari and Nwamae (2019). Principal Component Analysis (PCA) was introduced to address this problem for linear models. PCA identifies the eigenvectors $\mathbf{v}$ of the covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}\mathbf{v} = \lambda\mathbf{v} \tag{8}$$

then retain k components such that the explained variance ratio η satisfies:

$$\eta = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{m} \lambda_j} \geq 0.95 \tag{9}$$

All numerical features were first standardized using a standard scaler. PCA was then applied to compress the matrix. Figure 1 illustrates the explained variance ratio across the principal components.

Analysis: The cumulative variance curve (orange line) demonstrates that the first 16 principal components capture approximately 95% of the total variance from the original 31-feature set. This compression ratio suggests significant information redundancy in raw football metrics (e.g., the high correlation between shots on target and goals scored). By utilizing these 16 orthogonal components, the study effectively mitigates the "curse of dimensionality," providing a more stable input space for distance-based and linear algorithms like KNN and Logistic Regression.

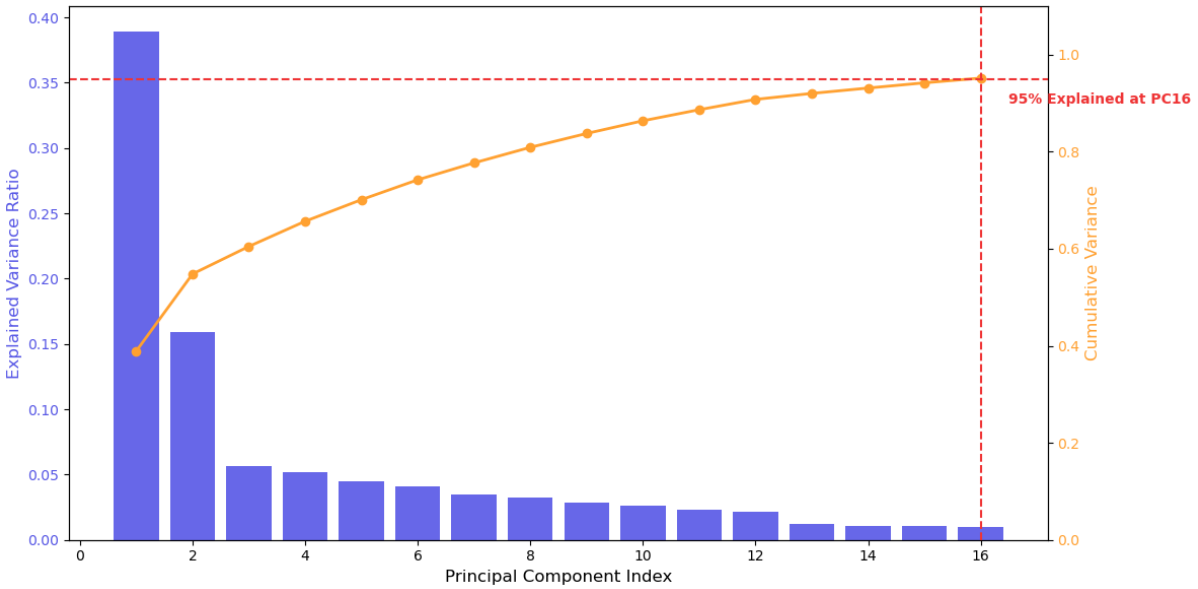

Figure 1: PCA Explained Variance Analysis

4 Predictive Modeling and Evaluation

After completing data preprocessing and feature engineering, the modeling and evaluation phase became the key step in validating the overall research approach. This section describes how the algorithms were selected, how model performance was measured, and how the final models performed relative to the benchmark set by the betting market.

4.1 Selection of Machine Learning Algorithms

At this stage, seven mainstream machine learning algorithms were introduced, each representing a distinct technical approach. Tree-based ensemble models like XGBoost and Random Forest were selected for their strong performance on nonlinear data. Logistic Regression, k-Nearest Neighbor, Artificial Neural Networks, and Naïve Bayes were included as reference models for comparison. Table 4 shows the input data selection methods for all models.

Within this set, CatBoost played a slightly different role. Conventional machine learning models may not interpret categorical names such as "Arsenal" or "Anthony Taylor" directly. Data scientists must convert such text into numerical form, often using one-hot encoding. CatBoost, by contrast, ingests raw categorical features directly, removing the need for manual encoding. When configuring CatBoost, the 15 numerical features selected by Recursive Feature Elimination were combined with raw text fields like team names and referee names. This hybrid feature matrix preserved the original meaning of categorical variables, making the model more sensitive to subtle influences from specific teams or referees.

Table 4: Feature Selection and Preprocessing Strategies by Model Architecture

Model	**Type**	**Feature Strategy**	**Preprocessing & Rationale**
XGBoost	Gradient Boosting	RFE Top 15	Label Encoding
Random Forest	Bagging Ensemble	RFE Top 15	Bootstrap aggregating
CatBoost	Ordered Boosting	RFE Top 15 + Categorical IDs (Team Name, Referee)	Native Handling
Logistic Regression	Linear	PCA	StandardScaler + PCA
KNN	Distance-based	PCA	StandardScaler + PCA
ANN	Neural Network	PCA	StandardScaler + PCA
Naive Bayes	Probabilistic	PCA	StandardScaler + PCA

4.2 Evaluation Metrics: Accuracy vs. RPS

Different fields tend to adopt different habits when evaluating model quality. In traditional machine learning tasks, accuracy is the most commonly used metric, calculated as the proportion of correct predictions among all predictions.

However, in betting, accuracy has a main limitation. Take two matches: in one, the model predicts a win probability of 51%; in the other, it predicts 90%. If both are correct, accuracy treats them the same. Yet, the first is barely better than a coin toss, while the second shows much greater confidence. These two predictions imply different risk and value, but accuracy reduces this to a simple binary result.

To address this issue, this study adopted the Ranked Probability Score (RPS) as the primary evaluation metric. RPS is widely used in sports prediction research and is designed to measure the distance between the predicted probability distribution and the actual outcome. The formula is as follows:

$$\text{RPS} = \frac{1}{r-1} \sum_{i=1}^{r-1} \left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j \right)^2 \tag{10}$$

where $r = 3$, p is the predicted probability vector, and e is the cumulative outcome indicator.

The calculation sums the squared differences between cumulative predicted probabilities and cumulative actual outcomes, then averages across the number of outcome categories.

A smaller RPS means the model's probability distribution is closer to the true distribution. RPS heavily penalizes confident but incorrect predictions. For example, if a model assigns a 95% probability to an outcome that does not happen, it receives a much higher error score. This quality makes RPS suitable for betting contexts, where overconfident mistakes often cause greater losses than uncertain ones. By using RPS, the evaluation shows both predictive accuracy and risk sensitivity.

4.3 Model Performance Comparison

Once the evaluation metrics were established, the predictive efficacy of the seven candidate algorithms could be assessed using the 2024–2025 season test set. Table 5 and Figure 2 present the synchronized results of Accuracy and Ranked Probability Score (RPS).

The results confirm that tree-based ensemble models (Random Forest, XGBoost) outperform linear and neural network approaches for sports forecasting. Random Forest achieved the highest accuracy (53.52%) and the lowest RPS (0.1997), verifying the theoretical advantage of "bagging" and "feature randomness" in handling the high-stochasticity environment of the English Premier League. While NaiveBayes shows a relatively high accuracy (52.67%), its RPS is inferior to that of Logistic Regression. This highlights the value of RPS: it penalizes Naive Bayes for potentially "overconfident" probability assignments, proving that a model can be right often but mathematically "clumsy" in its probability distribution. Crucially, the top models (RF, XGB) achieved an RPS below the market benchmark (approx. 0.2005), suggesting that algorithmic integration of Elo and form metrics can indeed find "alpha" or predictive edges not fully captured by bookmaker odds.

Table 5: Performance metrics across various machine learning models

Model	Accuracy	RPS
RandomForest	0.5352	0.1997
XGBoost	0.5309	0.2026
CatBoost	0.5309	0.2038
LogisticReg	0.5096	0.2069
NaiveBayes	0.5267	0.2074
KNN	0.5075	0.2222
ANN	0.4350	0.2862

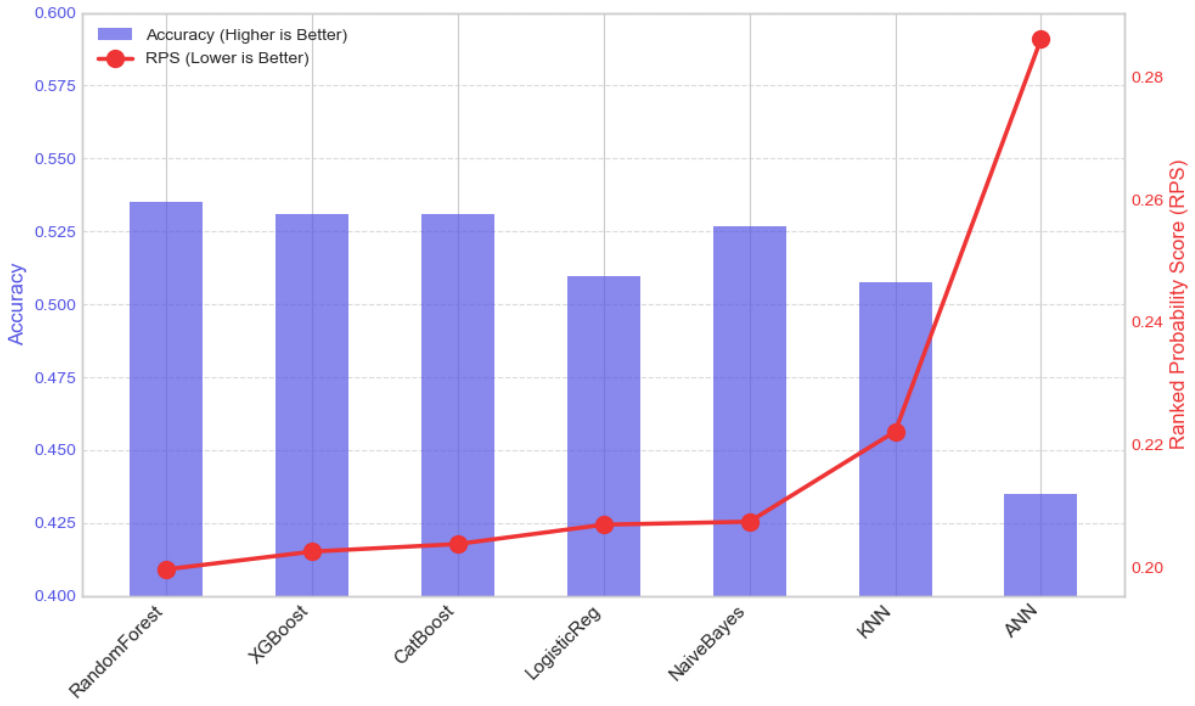

Figure 2: Prediction results from various machine learning algorithms

The calibration curve provides a visual assessment of the reliability of the predicted probabilities. As showed in Figure 3, the proximity of all four primary models (RandomForest, XGBoost, CatBoost, and NaiveBayes) to the ideal 45-degree diagonal indicates that the system is well-calibrated overall.

RandomForest (purple) and XGBoost (yellow) demonstrate high stability, particularly in the mid-to-high probability range (0.6 to 0.9). This implies that when these models assign a high win probability to a home team, the actual win rate closely mirrors that confidence level, providing a reliable foundation for value identification.

The NaiveBayes curve (dark red) shows more frequent oscillations and a distinct "S-shape" pattern in several segments. This indicates a tendency toward overconfidence or underconfidence at different intervals, which mathematically explains why it suffers from a higher RPS despite competitive Accuracy.

All models show increased noise at the lower end of the probability spectrum (below 0.3). This empirical evidence supports the study's decision to adopt Logic A (Most Likely Outcome). By focusing on events where the model is most confident and well-calibrated, the strategy avoids the "tail-end" calibration errors that plague more aggressive strategies like Logic B.Validation of Market Edge: The high alignment with the diagonal across multiple algorithms suggests that the feature engineering process, particularly the integration of Elo ratings and market implied probabilities, has successfully captured the underlying objective distribution of EPL match outcomes.

4.4 Feature Importance Analysis

Algorithms such as XGBoost are often described as "black boxes" because their internal decision-making processes are not immediately transparent. To deconstruct the "black box" of the best-performing model, feature importance weights were extracted as shown in Figure 4.

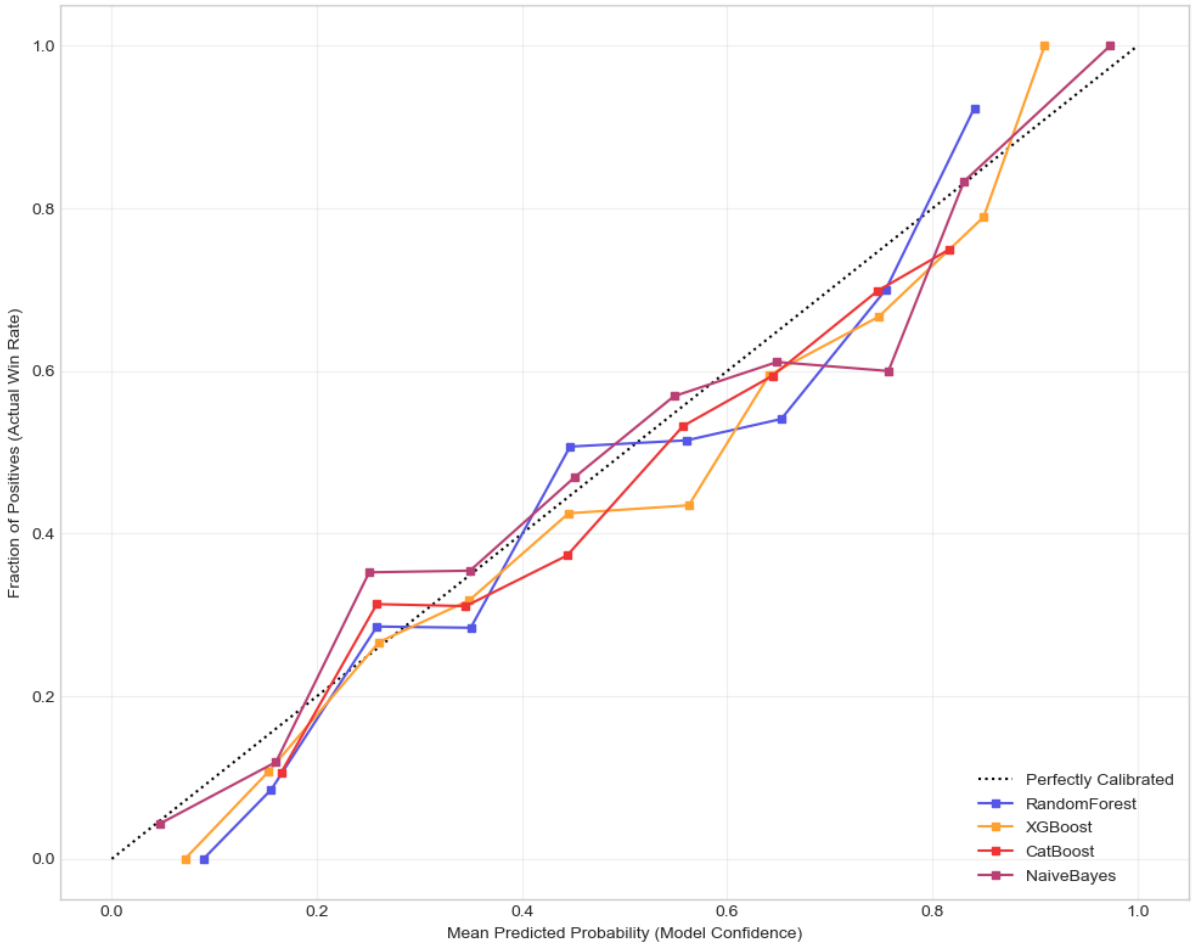

Figure 3: Calibration Curve: Predicted vs. Actual Win Rate

The ranking of feature importance revealed that market implied probabilities—such as ImpPH (implied probability of a home win) and ImpPA (implied probability of an away win)—dominated all other features. This finding is not surprising: bookmaker odds reflect the convergence of large-scale market participation and information, effectively serving as a collective judgment difficult to surpass.

Nevertheless, the high ranking of the Elo differential (EloDiff) and the Elo of away-team-win (EloAwayW) suggests that the Elo rating system provides a superior measure of "true underlying strength" compared to simple win-loss records.

Tactical Granularity: The inclusion of average shots differential (AvgSDiff) and the home team's defense and offensive ability (HomeDef/Off) among the top-tier features indicates that the model efficiently captured "process-oriented" metrics. This supports the academic argument that shot-based data is more predictive of future performance than goal-based data due to a larger sample size and lower variance.

4.5 Market Benchmark Testing

Throughout the study, bookmaker odds were a natural benchmark. Comparing models was a first step. The more meaningful challenge was to see if the models could approach or beat this market benchmark.

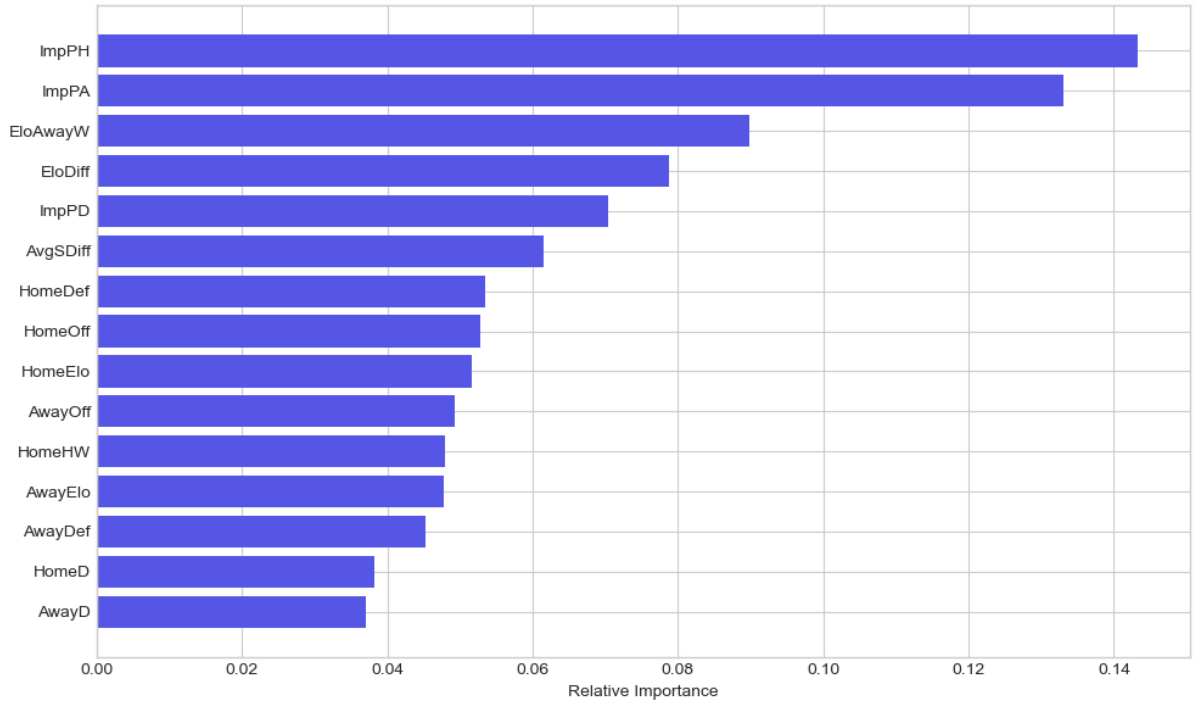

Figure 4: Feature importance in Random Forests model

To make this comparison, the original bookmaker odds from the test set were first extracted. The margin was removed to find the market's true consensus probabilities. These probabilities were used to calculate a baseline RPS score for betting, which was around 0.2005.

The RPS score of the Random Forests model (0.1997) was then directly compared against this market baseline. The result showed that the model achieved a slightly lower RPS, indicating a marginally better probability estimation.

In a domain as stochastic as sports betting, a difference of 0.008 may appear modest. However, at the level of probabilistic prediction, it suggests that the model estimated outcome probabilities with slightly greater accuracy than the collective judgment embedded in market odds. This finding delivers a quantitative foundation for the strategy formulation and capital management experiments in the subsequent sections.

5 Formulation of the Betting Strategy

Once the machine learning models demonstrated the ability to approach—and in some cases surpass—the market benchmark, the focus shifted to strategy design, the core of quantitative betting. While an accurately calibrated probability estimate is the prerequisite, translating these probabilities into consistent financial returns requires precise mathematical definitions and robust decision rules.

5.1 Investment Performance Metrics

To objectively evaluate the economic efficiency and risk exposure of the proposed betting strategies, two pivotal financial indicators are introduced: Return on Investment (ROI) and Maximum Drawdown (MDD). While Accuracy and RPS measure predictive precision, these metrics quantify the actual investment ability of the system.

5.1.1 Return on Investment (ROI)

ROI measures the efficiency of capital utilization by calculating the net profit relative to the total volume of capital wagered. In a betting context, it is defined as:

$$ROI = \frac{\sum_{i=1}^{n}(Profit_i)}{\sum_{i=1}^{n}(Stake_i)} \times 100\% \tag{11}$$

where n is the total number of bets.

An ROI of 10% implies that for every 100 wagered, the system generates a net profit of 10. In the high-efficiency English Premier League market, any consistent positive ROI represents a significant predictive edge over the bookmaker's margin.

5.1.2 Maximum Drawdown (MDD)

While ROI focuses on reward, MDD is the primary measure of downside risk. It identifies the largest peak-to-trough decline in the account bankroll during a specific period. The formula is:

$$MDD = \frac{Value_{peak} - Value_{trough}}{Value_{peak}} \times 100\% \tag{12}$$

MDD is a critical psychological and financial threshold. For instance, Strategy 1's MDD of -15.2% indicates that an investor would have faced a 1,520 loss at the lowest point of a losing streak. In quantitative sports betting, a lower MDD (such as Strategy 4's -7.3%) is often prioritized over higher absolute profit, as it demonstrates the strategy's resilience against the inherent stochasticity (randomness) of football outcomes.

5.2 Defining Betting Value

In quantitative betting, a "value opportunity" exists when the odds offered by a bookmaker are higher than the reciprocal of the true probability of an outcome. This study evaluates three common mathematical formulations used to quantify this value:

Expected Return (Edge).

$$Edge = P_{\text{model}} \times O_{\text{book}} - 1 \tag{13}$$

where P_{model} is the probability predicted by the model, and O_{book} is the decimal odds offered by the bookmaker.

Odds Ratio (Ratio).

$$Ratio = \frac{O_{\text{book}}}{O_{\text{model}}} - 1 \tag{14}$$

where O_{book} is the bookmaker's odds, and O_{model} is the fair odds implied by the model.

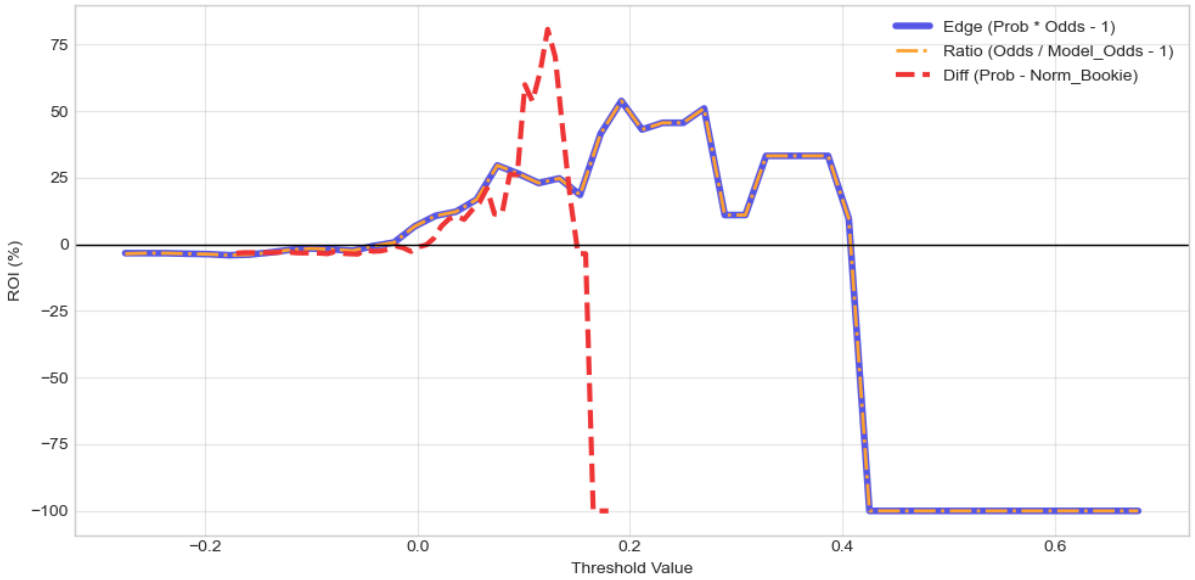

Figure 5: ROI Sensitivity Analysis across Value Thresholds

Absolute Probability Difference (Diff).

$$Diff = P_{\text{model}} - P_{\text{book}} \tag{15}$$

where P_{model} is the model-predicted probability, and P_{book} is the probability implied by the bookmaker's odds.

To evaluate the practical effectiveness of these formulations, a comparative experiment was conducted on the test set. Figure 5 illustrates the Return on Investment (ROI) profiles for each metric across a spectrum of value thresholds.

The empirical results reveal two critical insights. First, the Edge and Ratio formulations produce identical curves, confirming their mathematical equivalence; practitioners may use either interchangeably. Second, a prominent "variance trap" appears as the threshold exceeds 0.06. While ROIs initially peak, they subsequently oscillate sharply between +50% and −100%. This volatility is a function of sample size decay: as the system demands an increasingly extreme value advantage, the number of qualifying matches drops precipitously, allowing idiosyncratic randomness to overwhelm the model's predictive signal.

Furthermore, the Edge and Ratio metrics are susceptible to the favorite-longshot bias. A marginal 1% estimation error by the model on a favorite ($O = 1.5$) results in a negligible 1.5% spurious Edge. However, the same 1% error on a longshot ($O = 30.0$) is magnified into a 30% spurious Edge. To mitigate this "leverage risk" and maintain consistency with the RPS metric, the Absolute Probability Difference (Diff) was adopted as the primary value filter for all subsequent experiments.

5.3 Decision Logic

With Diff established as the value metric, the system must determine which outcome to select when multiple results (Home, Draw, Away) present positive value. Two distinct decision logics were back-tested using a fixed-stake ("Flat Betting") approach to isolate the impact of the decision rule from capital management variables.

Logic A (Most Likely Outcome with Value Filter) : The system identifies the outcome with the highest predicted probability (P_{max}) and executes the bet only if its $Diff$ exceeds a predefined threshold.

Logic B (Maximum Expected Value) : The system selects the outcome with the largest absolute $Diff$, regardless of its raw probability.The performance comparison of these two mechanisms is summarized in Table 6.

The core metrics used for evaluation are defined as follows:

- Bets: The total number of trades executed, representing market activity.
- Avg Odds: The average decimal price of the wagers, reflecting the favorite longshot orientation.
- Win Rate: The percentage of successful wagers.
- Max Drawdown (MDD): The largest peak-to-trough decline in capital, measuring tail risk.
- Return on Investment (ROI): Profit per dollar wagered, measuring the purity of the edge.

Table 6: Comparison of Decision Logics (Threshold = 0.02, Flat Betting)

Decision Logic	Bets	Avg Odds	Win Rate(%)	MDD(%)	ROI(%)
Logic A: Most Likely + Value	236	1.93	57.2	−15.2	6.3
Logic B: Maximum Value	389	3.15	43.2	−29.7	0.8

The results demonstrate a clear superiority for Logic A. Despite Logic B seeking the "maximum theoretical edge," it yielded a marginal ROI of only 0.8% and suffered a severe maximum drawdown of −29.7%. This underperformance stems from tail-end calibration errors.

Logic B systematically gravitates toward extreme underdogs (average odds of 3.15) because these outcomes often present the largest mathematical $Diff$ due to model instability at low-probability intervals. In contrast, Logic A functions as a two-stage defensive filter. By prioritizing the most likely outcome, it anchors the strategy to events with high frequency (57.2% win rate) and moderate odds (1.93), where the model's calibration is most reliable. The subsequent $Diff$ filter then ensures that capital is only deployed when a measurable margin of safety exists.

The empirical evidence suggests that in professional football forecasting, predictability at the core of the distribution is more valuable than theoretical value at the tails. Logic A provides a more robust foundation for sustainable capital growth by balancing hit rate with value margin.

6 Empirical Results

Building on earlier findings, Logic A (the most likely outcome with a value filter) was chosen as the base for the quantitative betting system. This section implements real-world risk constraints. It does so through ablation experiments, seeking the right balance between risk and return. The optimal strategy is then integrated into a dynamic capital management engine. This allows for assessment of the compounding effects of the Kelly Criterion.

6.1 Hypothesis-Driven Strategy Experiment

In quantitative finance, absolute return is often a secondary metric to risk-adjusted performance. To isolate the impact of specific business logic filters, five ablation strategies were tested within a flat-betting framework (fixed 100 per wager). Table 7 and Figure 6 show the result.

Strategy 1 (Naive Value) appears to be the most productive, yielding the highest Final Cap (11,479). This is primarily due to its high volume of Bets (236) and the highest Win Rate (57.2%). Because it targets favorites (with the lowest Avg Odds of 1.93), it wins frequently, but its ROI is modest at 6.3%. Critically, it is not an "optimal" strategy because its MDD (-15.2%) is the second highest. The high turnover amplifies exposure to bookmaker margins and variance, suggesting that its high profits result from "brute force" volume rather than a high-quality edge.

The most significant finding is the performance of Strategy 4 (Anti-Home Bias). By systematically excluding home-win bets, the Avg Odds rose to 2.17, while the MDD was slashed by more than half to -7.3%. Despite a lower absolute Final Cap due to reduced volume (86 bets), its 9.9% ROI is the highest in the study. This suggests that the betting market suffers from a "Home Bias" where casual bettors overvalue home advantage, leading bookmakers to shorten home odds. Strategy 4 exploits this inefficiency, providing the best risk-adjusted return.

Strategy 5 (Golden Spot) improved on Strategy 4 by capping odds at 3.0. This led to lower Win Rate (50.0%) and ROI (8.2%). MDD worsened to -11.0%. "Over-filtering" removed high-value away wins and shrank the sample size. That hurt the system's ability to manage losing streaks. This shows that robust logic (like anti-home bias) is better than strict numerical constraints.

Table 7: Detailed Data Comparison of Various Strategies

Strategy	Bets	Avg Odds	Win Rate(%)	MDD(%)	ROI(%)	Final Cap
Strategy 1: Naive Value	236	1.93	57.2	-15.2	6.3	11479.00
Strategy 2: Cap Longshots	227	1.87	58.1	-16.9	6.0	11364.00
Strategy 3: Mid-Odds Value	174	2.02	53.4	-14.4	7.8	11358.00
Strategy 4: Anti-Home Bias	86	2.17	53.5	-7.3	9.9	10853.00
Strategy 5: Golden Spot	68	2.15	50.0	-11.0	8.2	10557.00

6.2 Advanced Bankroll Management

Following the identification of Strategy 4 as the optimal value-capture engine, the study transitioned from static staking to dynamic capital management using the Fractional Kelly Criterion.

The Kelly fraction f^* is defined as:

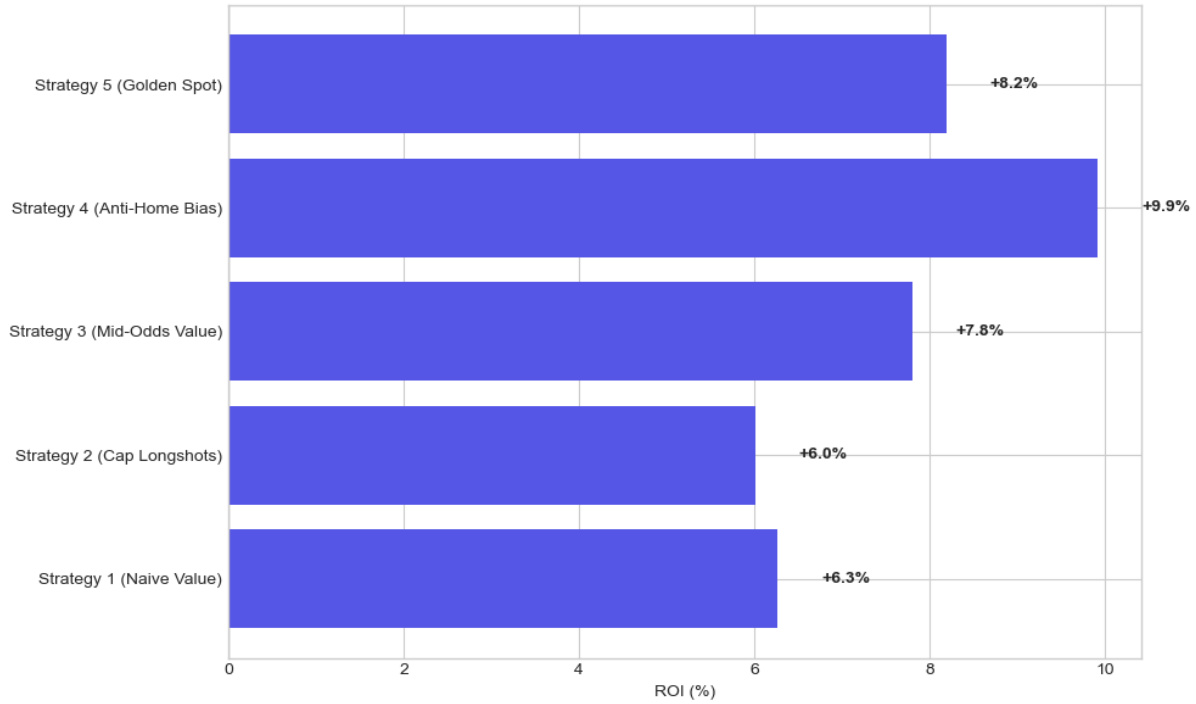

Figure 6: Performance of Hypothesis-Driven Strategies

$$f^* = \frac{b \times p - q}{b} \tag{16}$$

Here, b is the net decimal odds $(O - 1)$, p is the model's predicted probability, and $q = 1 - p$. This study used a conservative Quarter-Kelly (1/4). A 5% maximum cap per wager was set to protect the bankroll.

As Table 8 and Figure 7 show that the implementation of the Kelly Criterion yielded two profound improvements over flat betting.

Implicit Quality Filtering: The Kelly Criterion acted as a second filter. Flat betting had 86 matches, but Kelly rejected five with low expected value. This increased the win rate from 53.5% to 56.8% and lifted ROI to 15.3%.

Compounding Growth: The Kelly strategy reached a Final Cap of 12,787.93, tripling profit over flat betting (2,787 vs 853). Figure 6 shows the nonlinear growth trend. Max DD rose slightly (-10.3% vs -7.3%), but the trade-off is favorable. The system accepted 3% more drawdown for a 227% profit increase.

Table 8: Detailed Data Comparison of Staking Management

Staking Plan	Bets	Avg Odds	Win Rate(%)	MDD(%)	ROI(%)	Final Cap
Flat Betting (Fixed 100)	86	2.17	53.5	-7.3	9.9	10853.00
Fractional Kelly (1/4 Kelly)	81	2.21	56.8	-10.3	15.3	12787.93

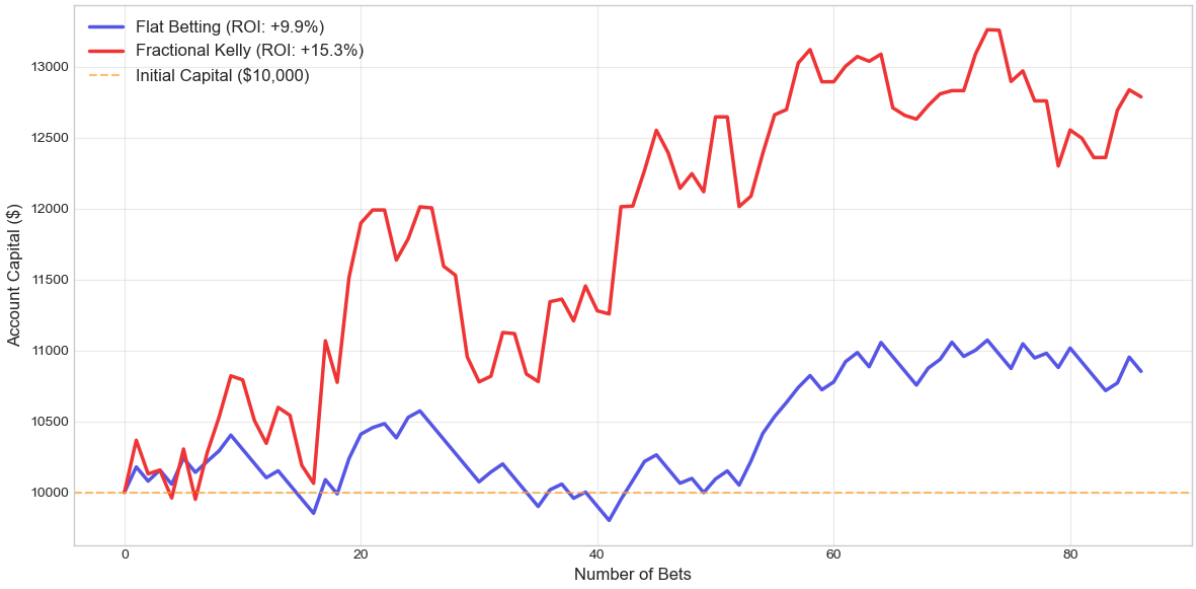

Figure 7: Bankroll Growth Comparison (Flat vs. Kelly)

By prioritizing high-edge opportunities with larger stakes while preserving capital during high-uncertainty periods, the Kelly-managed system transforms a moderate predictive edge into a high-performance investment vehicle.

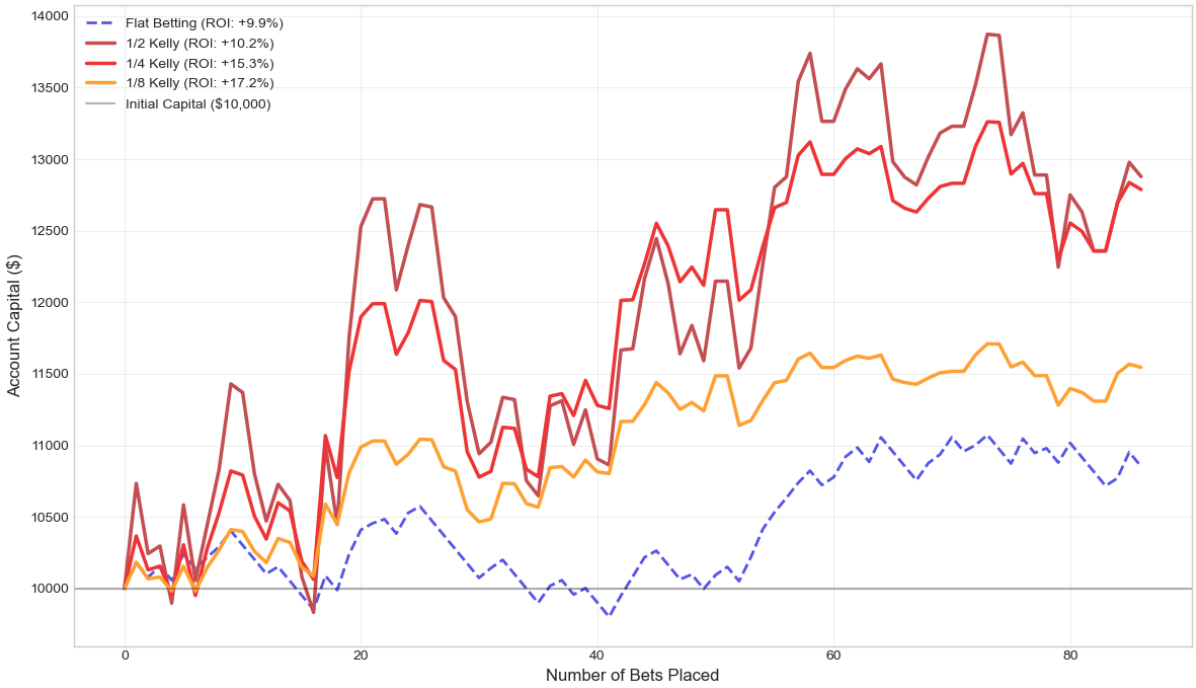

Figure 8: Kelly Fraction Sensitivity Analysis

To further scrutinize the robustness of the capital management engine, a sensitivity analysis was conducted by varying the Kelly fraction (f^*) across four levels: Flat Betting, 1/8 Kelly, 1/4 Kelly, and 1/2 Kelly. The results, shown in Figure 8, reveal the non-linear relationship between leverage and bankroll volatility.

The 1/2 Kelly strategy (dark red line) achieved the highest peak capital, demonstrating the potent compounding effect of the Kelly Criterion when the model maintains a persistent edge.

However, its trajectory is marked by extreme verticality in both directions, indicating high sensitivity to individual match outcomes.

While the 1/2 Kelly strategy reached the highest absolute value at several points, it also experienced the most severe drawdowns (e.g., around the 40th bet). This empirical evidence supports the theoretical caution that higher Kelly fractions, while maximizing logarithmic growth, significantly increase the probability of a "Gambler's Ruin" scenario during sustained losing streaks.

The 1/4 Kelly strategy (bright red line) presents the most favourable risk-adjusted profile. It captures a significant portion of Kelly's 1/2 upside while continuing a much smoother equity curve, particularly during mid-season volatility. The transition from 1/8 Kelly to 1/4 Kelly yields a considerable leap in Final Capital without a proportional increase in Maximum Drawdown.

All Kelly variants consistently outperformed Flat Betting (blue dashed line). This difference confirms that the model's predicted "Edge" is not only directionally correct but also well-calibrated; otherwise, dynamic staking would have accelerated losses rather than gains.

The sensitivity analysis supports selecting the Quarter-Kelly (1/4) approach for the final system. It strikes an ideal balance between the aggressive wealth-building properties of the Kelly index and the practical necessity of capital preservation in the stochastic environment of the English Premier League.

7 Walk-Forward Validation and Robustness Analysis

7.1 From Static Backtesting to Walk-Forward Validation

While initial backtesting provided a preliminary "proof of concept" for the predictive models, static evaluation often falls prey to overfitting and look-ahead bias. To simulate a realistic investment environment, this study implements a Walk-Forward Validation framework.

The core logic of this framework involves a dynamic "Expanding Window" approach: the models are initially trained on data from 2010 to 2017 to predict the 2018 season. Subsequently, the window expands to include 2018 data for retraining before predicting 2019. This iterative process ensures that the model adapts to Regime Shifts—systemic changes in league dynamics such as tactical evolutions, VAR implementation, or shifts in home-field advantage—without violating temporal causality.

Beyond the temporal restructuring of the dataset, the experimental design of the Walk-Forward Validation maintains strict consistency in model selection and strategy filtering to ensure causal clarity.

Model Selection (Random Forest): We standardized on the Random Forest architecture for each retraining cycle. This choice was predicated on its inherent resistance to overfitting and its ability to handle non-linear feature interactions without extensive hyperparameter tuning per cycle. By coupling RF with Recursive Feature Elimination (RFE), the model autonomously re-identifies the most influential variables (e.g., shifts in weighting between Elo and ODM) for each specific season.

Decision Logic (Logic A): While Logic B (Value Maximization) offers higher theoretical upside, Logic A (Maximum Likelihood) was selected as the primary engine. This approach prioritizes the "purity" of the predictive edge, focusing on the model's ability to accurately estimate the most probable outcome. This serves as a more rigorous test of the model's fundamental calibration across shifting market regimes.

Strategy Fixation (Strategy 4 - Anti-Home Bias): To test the universality of the "Home-Field Premium" hypothesis, Strategy 4 was held constant across all seven test windows. By maintaining a fixed filtering threshold ($Diff \geq 0.02$) and excluding home-win predictions, the experiment provides a longitudinal stress test of the contrarian edge. This consistency is vital for identifying the specific market conditions (regime shifts) under which the strategy either thrives or encounters systemic friction.

7.2 Performance Analysis

As shown in Figure 9, the rolling validation reveals a non-linear equity trajectory that contrasts sharply with the idealized growth often observed in static tests.

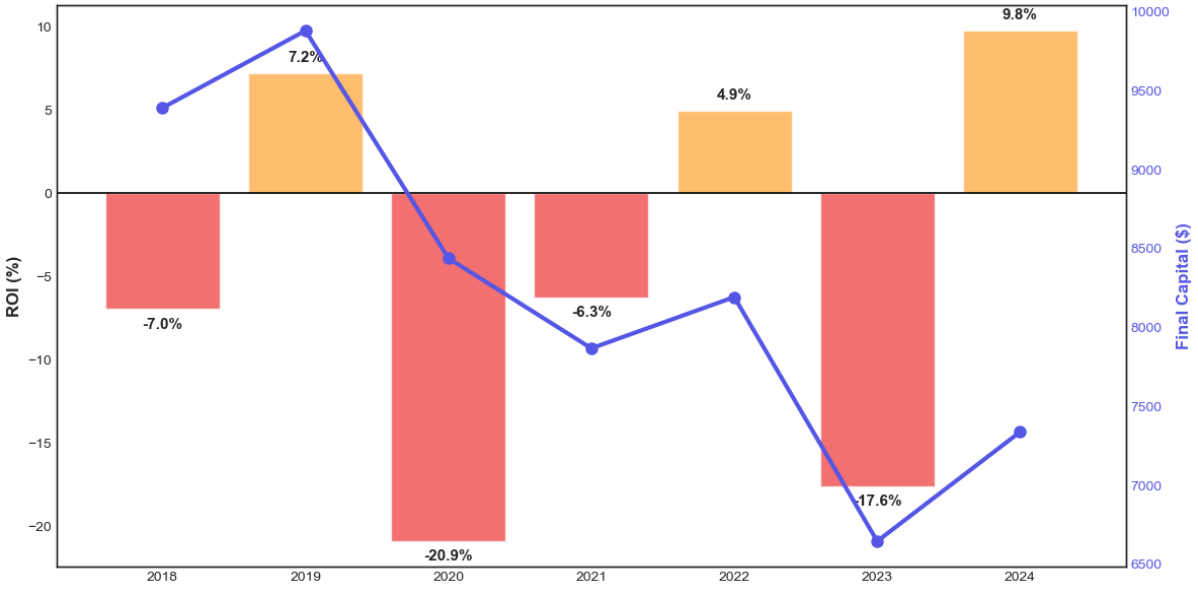

Figure 9: Performance under Walk-Forward Validation

The 2024 Recovery: A pivotal finding is the robust performance in the 2024 season, which achieved a 9.8% ROI and a 52.1% Win Rate. This rebound demonstrates the efficacy of the dynamic retraining mechanism: by absorbing the volatile 2023 data, the model recalibrated its feature weights to successfully identify alpha in the current market regime.

Drawdown and Resilience: The strategy encountered significant headwinds in 2020 and 2023, with ROIs of -20.9% and -17.6% respectively. However, preserving capital above a critical ruin threshold validates the risk-mitigating properties of Strategy 4 (Anti-Home Bias), which focuses on contrarian value rather than chasing market favorites.

7.3 Diagnostic Experiments

To move beyond a purely descriptive account of profit and loss, two diagnostic experiments were designed to decouple internal model calibration from external market dynamics. The goal is to provide a granular explanation for the volatility observed in the Walk-Forward Validation, particularly the drawdowns in 2020 and 2023.

7.3.1 Experiment A: Market Regime and Outcome Distribution

The primary objective of Experiment A is to test the hypothesis of Market Regime Shift. Since the core of our investment logic is Strategy 4 (Anti-Home Bias), the strategy's edge is theoretically contingent on the historical distribution of Home, Draw, and Away wins. By quantifying the actual outcome frequencies across the rolling test period (2018–2024), we aim to determine whether periods of underperformance were caused by a systemic surge in home-field advantage that deviated from historical norms.

As illustrated in Figure 10, the longitudinal distribution of match outcomes confirms a significant environmental shift.

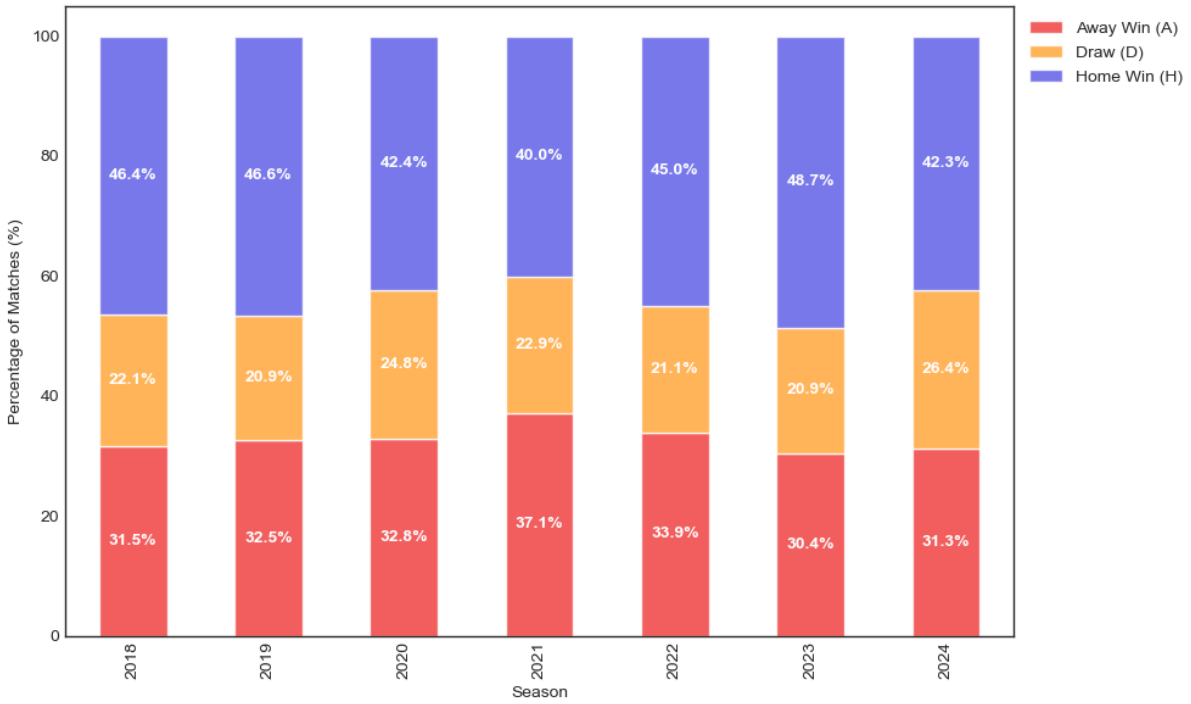

Figure 10: Actual Match Outcome Distribution per Season

In the 2023 season—the period of our most severe drawdown—the Home Win (H) rate surged to 48.7%, a historical peak that is nearly 7% higher than the 2021 and 2024 levels.

This surge indicates a "Regime Shift" where mid-table teams exhibited an unusually dominant home-field presence. For an Anti-Home Bias strategy, this environment represents a "tail-risk" event where market favorites consistently outperformed probabilistic expectations. This external evidence provides a robust defense for the model: the losses were not a failure of predictive logic but a result of a highly skewed market regime that neutralized the contrarian edge.

7.3.2 Experiment B: Confidence Segmentation and Calibration

While Experiment A addresses external factors, Experiment B investigates Internal Model Integrity. By segmenting the predictions into three confidence bins—Low (0.3–0.45), Mid (0.45–0.6), and High (0.6–1.0)—we aim to evaluate the model's Calibration Curve. The objective is to determine whether the model remains "rational" at high-probability levels or suffers from Overconfidence Bias, which often leads to disproportionate losses during volatile seasons.

The ROI distribution across confidence segments in Figure 11 reveals a nuanced internal diagnostic:

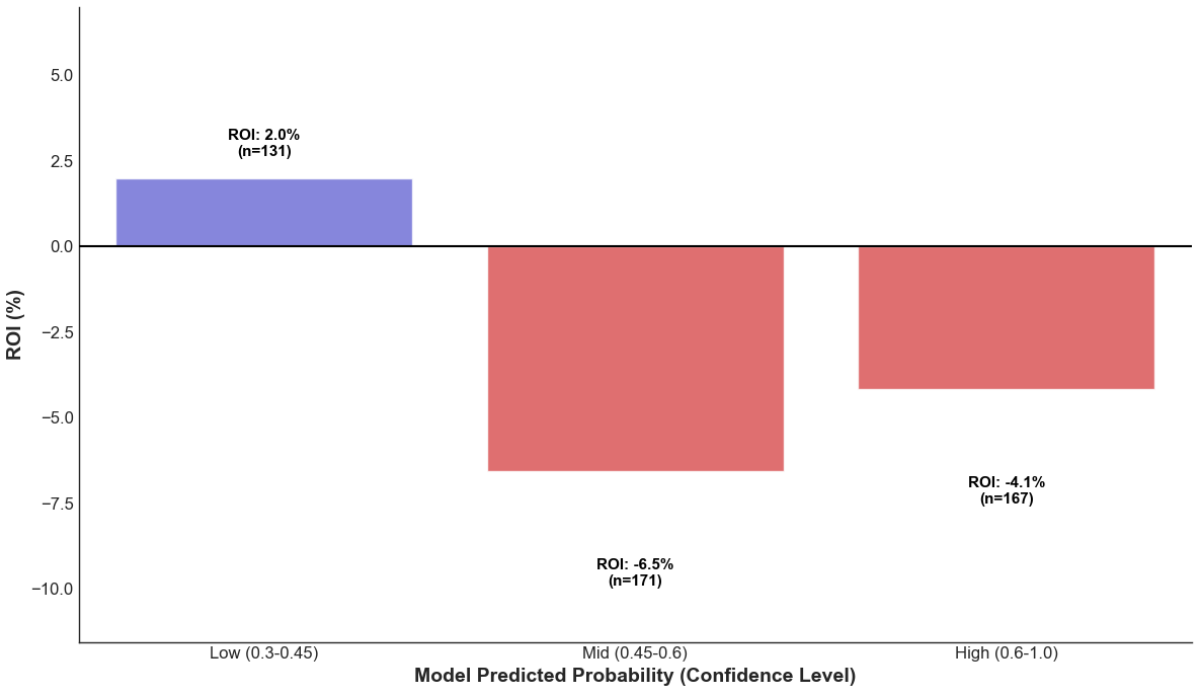

Figure 11: Profitability by Model Confidence Segment

The negative ROI (-4.1%, $n = 167$) in the high-confidence bin suggests a calibration drift. This indicates that when the model is most certain (Prob ¿ 0.6), it tends to underestimate the impact of stochastic variables (e.g., in-match red cards or sudden injury news), resulting in a failure to capture the market margin.

Interestingly, the Low (0.3–0.45) segment maintained a positive ROI of $+2.0\%$ ($n = 131$). This confirms that the model's true strength lies in identifying mispriced underdogs—situations where the collective market (Avg Odds) has heavily discounted a non-home outcome.

This performance validates the decision to use Strategy 4. It proves that the system's value is derived from its ability to act as a contrarian filter, successfully capturing "Longshot Alpha" even when its "Steady Favorite" predictions are challenged by market noise.

8 Conclusion

8.1 Summary of Key Findings

This study constructed a quantitative framework for sports betting. It analysed 5,790 English Premier League matches from 15 seasons. Three main findings emerged from this complete quantitative pipeline.

First, machine learning models can capture the complex tactical patterns embedded in football data. During the prediction evaluation stage, XGBoost and Random Forest performed best among the models tested. Their Ranked Probability Scores (RPS) approached the market benchmark of 0.2005, indicating that the models estimated match outcome probabilities with similar accuracy to the aggregated market.

Second, the gap between prediction and profitability depends on the decision mechanism. The absolute probability difference (Diff) was the most appropriate metric to quantify betting value. Empirical results showed that Logic A worked well. It first identified the most likely outcome and then applied a value filter. This approach avoided calibration errors that often occur with low-probability events. The ablation experiments showed that the betting market systematically overvalues home teams. The anti-home bias strategy simply excluded home win bets. It achieved a 9.9% ROI and kept maximum drawdown at only –7.3%.

Third, dynamic capital management can reshape a betting strategy's risk–return profile. Introducing the fractional Kelly Criterion led to a substantial improvement. Beyond compounding those amplified absolute returns, the Kelly Criterion also served as an implicit secondary filter. It automatically rejected bets where the expected value did not overcome the bookmaker's margin. This mechanism pushed the system's ROI to 15.3%.

8.2 Limitations of the Study

Although these encouraging results, several limitations should be acknowledged when considering real-world applications.

First, liquidity constraints and account restrictions pose practical challenges that are not captured in a back-testing environment. As Wilkens (2021) notes in a study on tennis betting, bookmakers actively monitor bettors who show consistent profitability. Accounts identified as "smart money" usually face reduced maximum stake limits or may be closed entirely—a practice known in the industry as "gubbing." The back-testing framework in this study assumed frictionless, unlimited market access, a difficult-to-satisfy assumption in practice.

Second, the feature engineering in this study relied entirely on pre-match aggregated statistics. Football matches are highly dynamic; events such as red cards, unexpected injuries, or sudden weather changes can shift the probability of outcomes in real time. A static pre-match model, by definition, cannot account for these in-play developments.

8.3 Future Work

Future research could expand this framework by using richer data sources and adopting more flexible models.

One direction is to incorporate player-level optical tracking data. As Stübinger et al. (2019-12-19) demonstrates, integrating metrics such as distance covered and passing accuracy can improve match-level prediction. Future models could also include advanced metrics such as expected goals and passing network data. These may help the model distinguish genuine tactical effectiveness from noise in raw scorelines.

Finally, the use of deep learning architectures for multidimensional time-series modeling offers additional promise. The current framework relies on manually constructed rolling averages to represent recent team form, which inevitably loses information about the temporal order of events. As Berrar et al. (2024-10) emphasizes, a football league is a complex time series with interacting entities. Future studies could explore architectures such as LSTM networks or Transformer-based models, which are designed to capture long-term trends and short-term changes end-to-end without the need for manually engineered features.

Bibliography

Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35, 2019.

Ryan Beal, Stuart E Middleton, Timothy J Norman, and Sarvapali D Ramchurn. Combining machine learning and human experts to predict match outcomes in football: A baseline model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Daniel Berrar, Philippe Lopes, and Werner Dubitzky. Incorporating domain knowledge in machine learning for soccer outcome prediction. 2019.

Daniel Berrar, Philippe Lopes, and Werner Dubitzky. A data- and knowledge-driven framework for developing machine learning models to predict soccer match outcomes. *Machine Learning*, 113, 2024-10.

Joel Brooks, Matthew Kerr, and John Guttag. Using machine learning to draw inferences from pass location data in soccer. 2016.

Rory P. Bunker and Fadi Thabtah. A machine learning framework for sport result prediction. 2019.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Anthony C. Constantinou. Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 108(1), 2019-01.

databridgemarketresearch.com. Global sports betting market size, share, and trends analysis report – industry overview and forecast to 2033, 2025. URL `https://www.databridgemarketresearch.com/reports/global-sports-betting-market`.

Arpad E. Elo. *The rating of chessplayers, past and present*. Batsford, 1978.

Engin Esme and Mustafa Servet Kiran. Prediction of football match outcomes based on bookmaker odds by using k-nearest neighbor algorithm. *International Journal of Machine Learning and Computing*, 2018.

Football-Data.co.uk. Historical football data, 2025. URL `https://www.football-data.co.uk/`.

Ledisi Kabari and Believe B Nwamae. Principal component analysis (pca)-an effective tool in machine learning. *Int. J. Advanced Research in Computer Science and Software Engineering*, 2019.

John L Kelly. A new interpretation of information rate. *the bell system technical journal*, 1956.

Siem Jan Koopman and Rutger Lit. Forecasting football match results in national league competitions using score-driven time series models. *International Journal of Forecasting*, 35, 2019-04.

Yujiao Li and Yingjie Mu. Research and performance analysis of random forest-based feature selection algorithm in sports effectiveness evaluation. *Scientific Reports*, 2024.

Philippe LOPES, Daniel Berrar, Jesse Davis, and Werner Dubitzky. The 2017 soccer prediction challenge. 2017. Publisher: OSF.

Marios-Christos Malamatinos, Eleni Vrochidou, and George A. Papakostas. On predicting soccer outcomes in the greek league using machine learning. *Computers*, 11, 2022-08-31.

Purnachandra Mandadapu. The evolution of football betting- a machine learning approach to match outcome forecasting and bookmaker odds estimation, 2024.

Uhrín Matej, Šourek Gustav, Hubáček Ondřej, and Železnỳ Filip. Optimal sports betting strategies in practice: an experimental review. *IMA Journal of Management Mathematics*, 2021.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 2018.

Yiming Ren and Teo Susnjak. Predicting football match outcomes with eXplainable machine learning and the kelly index, 2022.

Pieter Robberechts and Jesse Davis. Forecasting the FIFA world cup – combining result- and goal-based team ability parameters. In *Machine Learning and Data Mining for Sports Analytics*. Springer International Publishing, 2019.

Johannes Stübinger and Julian Knoll. Beat the bookmaker – winning football bets with machine learning (best application paper). In *Artificial Intelligence XXXV*, volume 11311. Springer International Publishing, 2018.

Johannes Stübinger, Benedikt Mangold, and Julian Knoll. Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, 10, 2019-12-19.

Conor Walsh and Alok Joshi. Machine learning for sports betting: Should model selection be based on accuracy or calibration? *Machine Learning with Applications*, 2024.

Conor Walsh and Alok Joshi. Machine learning for sports betting: Should model selection be based on accuracy or calibration? *Machine Learning with Applications*, 16, 2024-06.

Sascha Wilkens. Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, 7, 2021.

Harry Zhang. Exploring conditions for the optimality of naive bayes. *International Journal of Pattern Recognition and Artificial Intelligence*, 2005.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

München 31.03.2026

Place, Date

Signature