Predicting Outcomes and Profits: Machine Learning and Betting Strategies in the English Premier League

Bachelor Thesis

Written by:

Alexander Myrup (160363) & Sebastian Ehrhardt (162497)

Supervising By:	Professor Daniel Hardt
Date of Submission:	12.05.2025
Total Characters:	109,075
Total Pages	55

CBS

Table of Contents

Abstract

This paper explores the effectiveness of combining machine learning algorithms and various betting strategies to achieve profitability in the context of English Premier League football betting. Data from five Premier League seasons was gathered, comprising 1,900 matches, to evaluate several machine learning models, including CatBoost, Random Forest, Logistic Regression, Naive Bayes, XGBoost, LightGBM, KNN, SVM, and a Voting ensemble classifier. Four distinct betting strategies; Flat Betting, Threshold Betting, Value Betting, and Kelly Criterion, were integrated to assess the potential for profitability when model predictions are applied using diverse strategies. Unlike existing literature that emphasizes prediction accuracy alone, this study bridges predictive analytics with betting strategies, addressing a clear gap in research regarding application of machine learning to football.

Comprehensive evaluation using historical match data revealed significant differences in predictive performance and profitability among the tested models and betting strategies. The ensemble Voting method consistently achieved strong results, with notable ROIs across Kelly Criterion (133.63%), Flat Betting (19.19%), Threshold Betting (15.91%), and Value Betting (21.32%). CatBoost and Random Forest also performed robustly across multiple strategies. In contrast, models such as Naive Bayes and KNN underperformed, reflecting challenges with probability calibration and adaptability to football's complex data dynamics.

The findings provide concrete evidence that machine learning, when combined with targeted betting strategies, has the potential to enhance profitability, though it also carries inherent risks. This research contributes valuable insights to the fields of sports betting analytics and machine learning applications.

Introduction

The sports betting industry has grown significantly in recent years, becoming a central element within the broader gambling industry. Driven by increased online accessibility, mobile platforms, and widespread fan engagement, sports betting now constitutes one of the most dynamic and economically significant areas of gambling, particularly in Europe (EU MARKET - EGBA, 2025). Online sports betting has transformed the betting landscape, offering bettors unprecedented convenience, extensive market variety, and immediate access to information and odds adjustments. As a result, the online sports betting segment has shown consistent growth, currently representing about 39% of Europe's total online gambling revenue according to the European Gaming and Betting Association (EU MARKET - EGBA, 2025).

As the most popular sport in the world with over 5 billion fans (*The Football Landscape – the Vision 2020-2023 - FIFA Publications*, n.d.), football is also the most prevalent sport for wagering, capturing a 50% share of betting volume globally (Federazione Italiana Giuoco Calcio *et al*., 2024). Football's global appeal, consistent media coverage, and year-round calendar of matches make it attractive to bettors and bookmakers alike. Major football events, such as the FIFA World Cup, UEFA Champions League and Premier League games, draw immense betting action, generating substantial betting turnover and highlighting football's leading role in the betting market.

Sports betting is an activity that has bettors place a wager on the outcome of a sporting event. The bettor placing the wager plays against the bookmaker, when one side wins, the other loses. Decimal odds present stakes in a straightforward decimal format, offering a quick way to gauge an event's probability. If the odds of Team A winning are 3.00, then a $1 wager would return $3 upon victory. If the bettor loses, only the wagered amount ($1) is lost. Higher odds, indicate higher returns, but also carry with them elevated risk (lower probability of occurring). Decimal odds also simplify odds comparison for bettors, eliminating the need for complex calculations and clearly indicating potential winnings, the higher the decimal, the greater the possible return. Bookmakers set odds carefully to ensure they make a profit, no matter who wins. This built-in

advantage for bookmakers is called the "*overround*" and usually makes it hard for bettors to consistently make money (Newall, 2015).

Football matches differ from many other sports in that they can end in three outcomes: a win, a loss, or a draw. A win occurs when one team scores more goals than their opponent, thus securing victory. Conversely, the team that scores fewer goals experiences a loss. A draw happens when both teams score an equal number of goals. The possibility of a draw adds complexity and uncertainty to football betting, making accurate outcome prediction particularly challenging and valuable. This also means that if choosing an outcome at random, there is a 33.3% chance of choosing the correct result.

Compared to traditional financial markets for equities, securities or derivatives, the sports betting market has the advantage that the results of a prediction become clear quickly. And should a prediction be correct or incorrect, payouts or losses are realized instantly without fluctuating results based on the overall market. This makes the sports betting market ideal for testing data-driven predictive models. However, having good predictions alone does not always guarantee profit. To succeed in sports betting, choosing the right betting strategy is equally important.

This paper will explore exactly how effective machine learning models are at generating sustainable profits in football betting, focusing on the English Premier League. This league was chosen due to its global popularity, abundant available data, and active betting market. By applying a wide range of ML models to this market and leveraging their predictions in conjunction with various betting strategies, this paper hopes to determine if a sustainable strategy can be created that yields an attractive return on investment (ROI).

The guiding research question that must be answered in this exploration is:

How effective are machine learning models in generating sustainable profits within the football betting market?

To answer this question, this study will apply advanced ML models, train and evaluate them using real Premier League matches from the last 5 seasons, and assess their performance using various betting strategies. Ultimately, this research aims to clarify whether using machine learning in football betting can lead to consistent and reliable profits.

Literature Review

In developing our research methodology, we primarily drew upon two foundational papers: *Data-driven prediction of soccer outcomes using enhanced machine and deep learning techniques*, (Mills et al., 2024) and *Beating the Bookmakers using Machine Learning* (Wijk, 2021). The first paper served as an essential methodological reference point, providing detailed guidance on using machine learning for football outcome prediction. It employed advanced machines and deep learning methods, though with notable differences to our approach. Specifically, their research focused on predicting both match results and the occurrence of more than two goals (over/under 2.5 goals) across the Dutch Eredivisie League, a league with noticeably different playing styles compared to the English Premier League. Unlike our study, they did not apply betting strategies to their predictions (Mills et al., 2024), presenting an opportunity to build on their foundational approach by incorporating financial outcome simulations and betting strategy assessments.

The second core paper, *Beating the Bookmakers using Machine Learning*, significantly influenced our emphasis on betting strategies and money management techniques. This paper's exploration of betting systems underscored the importance of aligning predictive accuracy with financial strategy to maximize return on investment. Their methodology explicitly considered various betting strategies, providing a robust framework for our own strategic evaluations, particularly regarding optimal stake-sizing and value identification (Wijk, 2021). The general consensus in this paper is that football betting markets are not always efficient, there are enough inefficiencies that can be exploited using machine learning, particularly by identifying and betting on mispriced odds.

A broader review of current literature (Stübinger et al., 2019; Yeung et al., 2024; Bunker et al., 2024) reveals numerous studies exploring the application of machine learning in predicting football match outcomes, reflecting growing academic and practical interest in predictive sports analytics. However, a noticeable gap persists regarding the integration of predictive models with betting strategies aimed explicitly at profitability rather than mere predictive accuracy. Most studies stop at prediction, with limited or no analysis of the practical profitability or financial viability executing these predictions in live betting markets. Consequently, our research aims to fill this void by evaluating machine learning predictions explicitly within a betting context, directly measuring success through return on investment (ROI) over a season and exploring how effectively these predictive insights translate into profitable financial outcomes.

Methodology

This chapter walks through every choice made to build, tune and assess the betting pipeline. It opens with a Philosophy of Science section that sets out the critical-realist, post-positivist ground rules that inform the rest of the study. Research Design follows, outlining the exploratory-confirmatory split and the single-season hold-out test. Next come the practical stages: Data Acquisition and Cleaning, where raw Premier League records are collected and sanitized; Feature Engineering, where rolling form metrics and team-level ratios are constructed; and Machine Learning, where eight candidate algorithms and an ensemble voter are specified. The chapter closes with the Evaluation Protocol, which details baselines, betting strategies and performance metrics, and with a short Ethical Considerations note on responsible use.

Philosophy of Science

Our project adopts a critical-realist stance in which match outcomes are material events while bookmaker odds are social artefacts that condense collective belief at a given moment. This layered reality guides a post-positivist epistemology, as every model is a provisional probe, and it’s worth is tested against fresh, unseen data rather than judged as final truth.

Knowledge is advanced through an abductive workflow with three key steps. First, deduction, efficient-market theory supplies the null expectation of zero systematic profit (Fama 1970). Second, induction, wide rolling windows let the data surface patterns that the efficient-market theory did not anticipate. Third, abduction proper, surprising signals prompt new features, revised hypotheses, and a return to deductive testing. The pipeline is frozen once exploration stabilizes, a single preregistered evaluation on the 2023-24 hold-out season serves as the present but not final verdict, since theoretically in a real-world application, each August the model will be retrained to include the completed season.

Axiological commitments are explicit, as return on investment is the headline metric because the research question asks whether an actionable edge exists, yet ROI is reported alongside Brier score and calibration curves so that financial gain is not mistaken for epistemic strength. We acknowledge that profitable deployment can feed back into price formation and amplify gambling harms. Results are therefore published with responsible-use guidance.

Methodological pluralism underpins the choice of an ensemble voting layer. Logistic regression, tree ensembles and kernel methods each highlight different facets of the game-market system, and their convergence offers a more resilient prediction than any single algorithm could provide.

Research Design

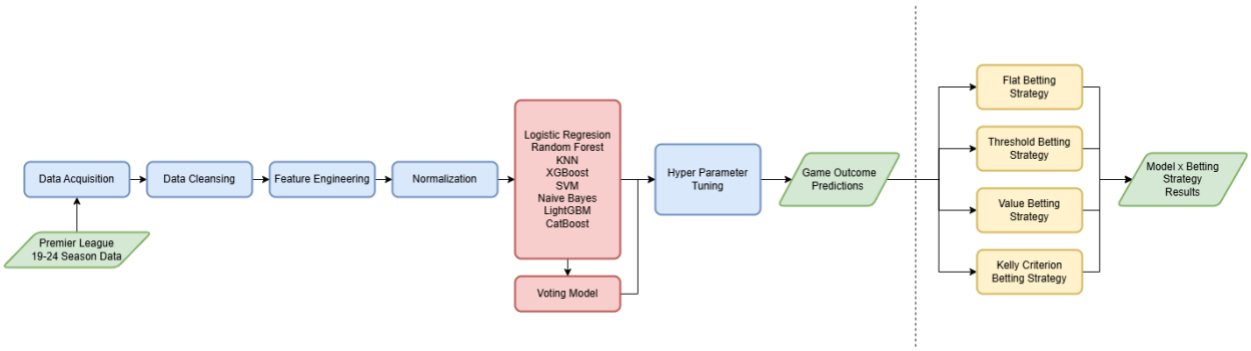


Figure 1: *Illustration of methodology.*

As detailed in figure 1, the process begins with acquiring historical data of matches played in the premier league from the 2019/20 season until and including the 2023/24 season. The data is then

cleansed, by dropping unnecessary columns, one hot encoding the teams playing and transforming the target to a categorical variable. New features are then created using a combination of rolling averages and team-level metrics, to improve the accuracy of the models. The data is then normalised to prevent features with larger scales from dominating. The data is then split into training and test data (80% of the data was used for the training dataset, 20% used for test dataset), retaining the chronological order to avoid lookahead bias, and the 8 models are trained and then tuned to have the highest return on investment on a validation set. Their predictions for each game are then stored, and a voting model then also makes a prediction based on the other model's predictions and it is added to the predictions. 4 different betting strategies are then applied to each model, each model is then evaluated on its return on investment across different betting strategies on the test data. Treating the 2023-24 season as a live falsification attempt operationalises our postpositivist claim that every result is provisional, any failure to profit there would immediately challenge the earlier exploratory findings.

Data

In this section we go through the processes of acquiring data, analysing it, creating features, and selecting them, and finally cleaning, normalising, and standardising the data. The quality and relevance of data are critical components in the performance of machine learning models, especially in sports analytics where data intricately reflects real-world scenarios and dynamic conditions. We carefully considered the characteristics unique to football, to ensure our features effectively captured meaningful relationships. Additionally, this section highlights our approach to handling missing values, mitigating biases, and transforming data into formats optimal for model training. By meticulously addressing these stages, we aimed to ensure the accuracy, reliability, and interpretability of our subsequent predictive modelling and betting strategy evaluations.

Data Acquisition

We were able to collect all the data we needed from one source (*England Football Results Betting Odds - Premiership Results & Betting Odds*, n.d.), which is incredibly convenient as it makes the data cleaning and preprocessing stage simpler, especially because the data quality was incredibly high. We considered other sources such as the official data partner of the Premier League, Oracle, however, although they had far more in-depth data and statistics and offered analytical tools, we decided to only use public data for replicability's sake. We also planned to use an API to include ELO ratings, which is a dynamic system commonly used in chess, that quantifies the relative skill levels of players (in this case football teams) in head-to-head competitions by updating scores based on game outcomes (Elo, 1978), however the API was incredibly unstable, and we again decided to exclude it again for ease of replicability. Calculating ELO Ratings manually was also deemed out of scope for this project. The only data quality challenge we had were certain stats only being collected during certain time periods, but this will be tackled later in the data cleaning stage.

When deciding on how large we wanted our dataset, our initial approach was the more the merrier, however we ended up settling on the last 5 complete seasons (Season 19/20 to 23/24) for two reasons. The first reason was that we wanted our data set size to reflect the dynamic nature of club's performances and the fast-paced evolution of football tactics and playing styles. By only including the last 5 seasons, we were able to limit the impact fundamental changes to the way football is played, which often leads to one or more teams far outperforming expectations. An argument can be made that this would be reflected in their stats, but for the scope of this project, we wanted to limit the time period to one where most tactical approaches have matured, and there have been no groundbreaking changes to managerial approaches in said period. The second reason is that we did not want historical performances of large clubs to impact the test set predictions. Limiting the horizon in this way emphasises our critical realist view that long-expired seasons reflect different underlying mechanisms and would blur today's causal structure. If we had used 10, 15, 20 years of data, we were concerned that our models would struggle to predict the outcomes of games for clubs that have experienced immense success in

the last 5 years, or a significant fall from grace. A potential opportunity for further research would be attempting to answer the same research question with a larger data set spanning across several years, exploring what statistical fundamentals have not changes over the last 30 years of premier league football.

Exploratory Data Analysis

The premier league football data file gives us the essentials: Date, HomeTeam, AwayTeam and a rich block of full-match statistics (shots, corners, fouls, yellow and red cards, goals) reported separately for each side, plus a few contextual fields such as Referee and kick-off Time. For exploratory purposes, these columns let us verify the sample size (1900 games) and confirm season coverage across all 5 seasons. Before modelling, however, we must distinguish what is known pre-kick-off from what appears afterwards to avoid data leakage. Doing so turns our ontological split between pre-match beliefs and post-match facts into a concrete data rule. Referee names and exact start times add little signal and will be dropped, as all bets will be placed simultaneously in our simulations, while pre-match identifiers will survive into feature engineering, where the performance statistics and market prices will be converted into rolling form indicators and historical-odds trends; once those transformations are created, the original match stats and odds columns are removed prior to model training.

Columns	Description
Date	Contains the date the game was played. Ranges from 09/08/2019 to 19/05/2024
Time	Contains the start time of the game in military time at Greenwich Mean Time. Ranges from 12:00 to 20:15
Referee	This contains the name of the referee in charge of officiating the game
HomeTeam	Contains the team playing at their home stadium
AwayTeam	Contains the team playing away from their home stadium

Table 1: *Match Information*

Columns	Description
FTHG	Full-time home team goals
FTAG	Full-time away team goals
FTR	Full-time result: H (Home win), D (Draw), or A (Away win)
HTHG	Half-time home team goals
HTAG	Half-time away team goals
HS	Home team shots (full-time)
AS	Away team shots (full-time)
HST	Home team shots on target
AST	Away team shots on target

Columns	Description
AF	Away team fouls
HC	Home team corners
AC	Away team corners
HY	Home team yellow cards
AY	Away team yellow cards
HR	Home team red cards
AR	Away team red cards
HF	Home team fouls

Table 2: *Match Statistics*

Betting Information

The file also includes 82 columns of pre-match odds supplied by six bookmakers; Bet365, Bet & Win, Interwetten, Pinnacle, William Hill, and VC Bet, which cover the home-draw-away line, over or under 2.5 goals, Asian handicaps, and corners. Because these prices are posted before kick-off they can legitimately feed the models, so at this exploratory stage we keep them in their raw decimal form. They give us a direct snapshot of the market's collective expectations and will later allow us to derive implied probabilities or value flags, but no rolling or drift features have been built from them, and nothing is dropped at this point.

Columns	Description
B365H/B365D/B365A	Bet365 home/draw/away win odds
BWH/BWD/BWA	Bet&Win home/draw/away win odds
IWH/IWD/IWA	Interwetten home/draw/away win odds
PSH/PSD/PSA	Pinnacle home/draw/away win odds
WHH/WHD/WHA	William Hill home/draw/away win odds
VCH/VCD/VCA	VC Bet home/draw/away win odds
MaxH/MaxD/MaxA	Market maximum home/draw/away win odds
AvgH/AvgD/AvgA	Market average home/draw/away win odds

B365>2.5/B365<2.5	Bet365 over/under 2.5 goals
P>2.5/P<2.5	Pinnacle over/under 2.5 goals
Max>2.5/Max<2.5	Market maximum over/under 2.5 goals
Avg>2.5/Avg<2.5	Market average over/under 2.5 goals
AHh	Market size of the Asian handicap (home team)
B365AHH/B365AHA	Bet365 Asian handicap home/away team odds
PAHH/PAHA	Pinnacle Asian handicap home/away team odds
MaxAHH/MaxAHA	Market maximum Asian handicap home/away team odds
AvgAHH/AvgAHA	Market average Asian handicap home/away team odds
B365CH/B365CD/B365CA	Bet365 corners home/draw/away win odds
BWCH/BWCD/BWCA	Bet&Win corners home/draw/away win odds
IWCH/IWCD/IWCA	Interwetten corners home/draw/away win odds
PSCH/PSCD/PSCA	Pinnacle corners home/draw/away win odds
WHCH/WHCD/WHCA	William Hill corners home/draw/away win odds
VCCH/VCCD/VCCA	VC Bet corners home/draw/away win odds
MaxCH/MaxCD/MaxCA	Market maximum corners home/draw/away win odds
AvgCH/AvgCD/AvgCA	Market average corners home/draw/away win odds
B365C>2.5/B365C<2.5	Bet365 corners over/under 2.5
PC>2.5/PC<2.5	Pinnacle corners over/under 2.5
MaxC>2.5/MaxC<2.5	Market maximum corners over/under 2.5
AvgC>2.5/AvgC<2.5	Market average corners over/under 2.5
AHCh	Market size of the corners Asian handicap (home team)
B365CAHH/B365CAHA	Bet365 corners Asian handicap home/away team odds
PCAHH/PCAHA	Pinnacle corners Asian handicap home/away team odds
MaxCAHH/MaxCAHA	Market maximum corners Asian handicap home/away team odds
AvgCAHH/AvgCAHA	Market average corners Asian handicap home/away team odds

Table 3: *Betting Information*

Visual explorations

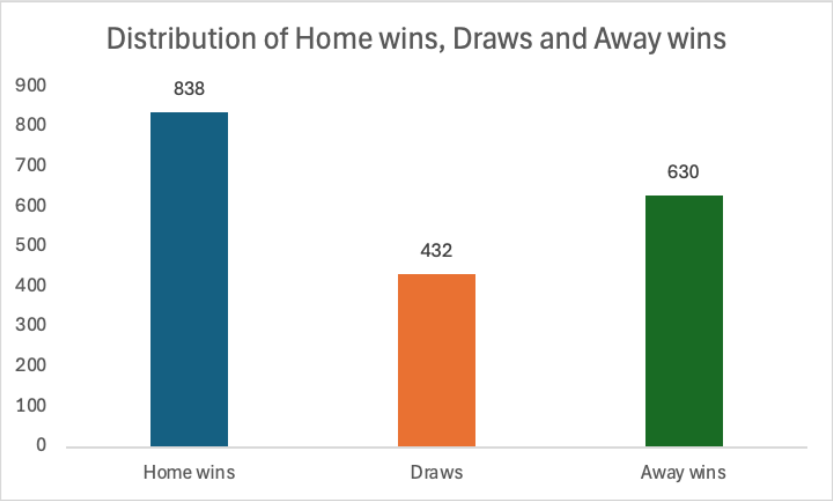


Figure 2: *Distribution of Home wins, Draws and Away wins.*

Figure 2 shows the split between home wins, draws and away wins in our dataset. This suggests that there is a class imbalance, with a significant home advantage, as teams tend to win more home than away. There are also far fewer draws than home wins and away wins. When doing feature engineering, we can keep this in mind in order to help the models distinguish between which team is home and which is away. This could be done through distinguishing between the home teams recent form and recent home form.

Another consideration derived from figure 2 is that we want to ensure that our models do not become prone to the accuracy paradox. The accuracy paradox is a term coined by Valverde-Albacete and Peláez-Moreno (2014) that describes the phenomenon that "*predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy*". If a model solely predicts home wins, it will have an accuracy of approximately 44%, which is better than a random guess which would have a 33% accuracy, however, would offer no real predictive insight. Therefore, it is important that we do not solely rely on accuracy as a metric of a good model.

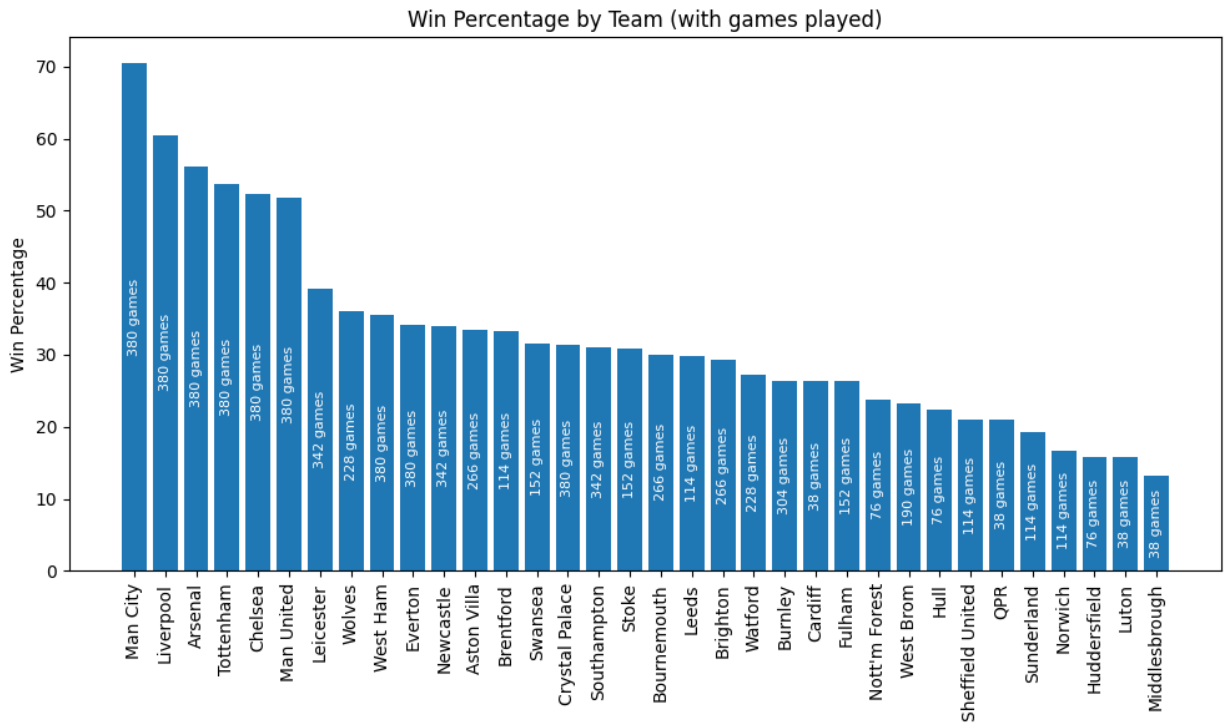


***Figure 3**: Teams' win percentages and games played from 19/20 to 23/24 seasons.*

Figure 3 shows that some teams, such as Man City, have a high win rate, while other teams such as Norwich, have a significantly lower win rate. This needs to be considered when creating features, as if the teams are encoded in the data, models may overfit certain teams and rather than relying on all the unique features. This would also be an issue as there is significant concept drift in our data. Concept drift refers to changes over time in the statistical relationship between input data and the target variable, meaning the model's underlying assumptions become outdated (Gama et al., 2014). An example would be that teams (even Man City) go through periods of strong form and weak form and our models should be able to handle these shifts in form. Another example (although outside our data's timeframe) would be Manchester United, having won 12 Premier League titles since 1992 (*Manchester United History*, n.d.). However, they are on track to record their worst finish in the Premier League in their history, with their second worse finish being the season prior (Bonn, 2025).

Outliers and anomalies

In this chosen dataset, there are no significant outlier results, however in football in general there are cases of teams severely outperforming expectations not only in a single game, but across a whole season. A notable example is Leicester City in the 2015/2016 season, who won the premier league despite beginning the season with the odds of winning being 5000 to 1 (*Leicester's Premier League Heroes*, n.d.). This presents both a challenge and an opportunity, as our models need to be adaptive enough to recognize a change in trends for a certain team, but this also creates an opportunity as if our models are able to spot the underdogs that are likely to overperform expectations, we can exploit it in our betting strategies in order to maximize our return on investment.

Feature Engineering

In this section, we outline the processes undertaken to create and refine features to enhance the predictive capability of our models by highlighting certain aspects within the data. These processes include encoding categorical variables, generating new metrics based on the stats available for games already played and computing rolling averages of the home and away teams' stats from previous games.

Encoding Categorical Variables

To effectively utilize team identifiers, we applied one-hot encoding to the *HomeTeam* and *AwayTeam* columns. This conversion created binary variables representing each team, facilitating their inclusion in predictive models without implying any unintended hierarchical relationships. We also considered using label encoding, assigning each team a number, and having a home and away column, however we were worried that giving the teams a numerical value would imply a relationship between the size of the number and the odds of a team winning/losing.

Team-Level Metrics

We calculated several metrics to encapsulate key performance indicators and improve predictive accuracy by capturing different dimensions of team performance. Goal Difference was computed separately for home and away teams to reflect scoring efficiency, directly correlating with offensive and defensive strength, as well as overall team form. Shot Accuracy Ratios, defined as the ratio of shots on target to total shots (*home_st_ratio* and *away_st_ratio*), were included to illustrate attacking precision and a team's ability to create high-quality scoring opportunities, under the assumption that better chances typically lead to more accurate shots. Additionally, Goal Conversion Rates were calculated as the number of goals scored relative to total shots, offering critical insight into a team's offensive effectiveness and their efficiency at converting opportunities into actual goals. Disciplinary Ratios, measured as the number of fouls committed per card received (*home_foul_to_card_ratio* and *away_foul_to_card_ratio*), provided insights into team discipline and aggression, potentially influencing match outcomes through suspensions, tactical fouling, or an ability to cope with physically aggressive opponents. Collectively, these metrics were strategically selected based on their direct influence on match dynamics and their potential to significantly enhance both model interpretability and predictive performance.

Rolling Averages

To account for recent team performance and form trends, which are vital when analysing a football team’s chance at success in a match, rolling averages of performance statistics were computed over intervals of 1, 3, 5, and 10 matches. Short-term intervals (1-3 matches) were included to capture immediate form fluctuations, while longer intervals (5-10 matches) provided insights into sustained performance trends. Metrics include goals scored, shots, corners, fouls, disciplinary actions as well as the previously mentioned ratios. By employing multiple intervals, our models gain nuanced insights into both temporary form fluctuations as well as consistent performance patterns. These rolling windows also enact the abductive loop because unexpected spikes trigger new hypotheses and feature revisions that we test against our inherent assumptions about market efficiency and baseline form indicators.

Limitations and Considerations

While our feature engineering approach significantly enhanced dataset quality, there were inherent limitations. Primarily, the features were constrained by publicly available data (by choice), excluding potentially influential factors such as player injuries, player form, detailed tactical formations, or psychological aspects like team morale. Future research could benefit from integrating these dimensions, via alternative data sources or more advanced analytical methods. Additionally, for further applications, we considered incorporating larger rolling window intervals and ELO ratings to capture broader performance trends and relative team strength. However, we deemed these out of scope for this project, given the already substantial volume of features included.

Final Feature Selection and Dataset Preparation

After introducing these new features, we removed redundant raw statistics to minimize multicollinearity and reduce noise. Additionally, the match outcomes (*FTR*) were encoded numerically as Home Win (0), Draw (1), and Away Win (2) to serve as our target variable. Column names were also standardized to ensure compatibility with machine learning algorithms. Through these steps, we improved the overall quality and predictive potential of our dataset, providing a solid foundation for subsequent model training and evaluation.

Data Cleaning and Preprocessing

Removal of Redundant Features

Initially, we identified and removed features that created data leakage, such as statistics that would only be available after a match has been played (e.g., final score, shots, corners, goal difference, etc.). This also included half time statistics, as they would not be available prior to the game. We then identified features that introduced noise, such as referee details and division labels (not relevant as all games are played in the same division). These features have no influence over the outcome of the game and therefore would not give the models any increase in accuracy. This step resulted in the removal of several columns, streamlining the dataset and

ensuring that our predictive models relied solely on information realistically available before match commencement.

Handling Missing Data

To manage missing data, we applied the dropna() function in pandas, opting for row removal rather than imputation. The decision to remove rows instead of imputing missing values was due to the minimal occurrence of incomplete data within our dataset. This choice ensured data consistency without the risk of introducing artificial biases that could result from data imputation, and follows post-positivist caution, avoiding extra modelling assumptions that might later be falsified.

Normalization and Column Standardization

We used standard scaling (StandardScaler) to normalize numerical features, ensuring that each feature contributed equally to model performance regardless of their original scale. Standard scaling was specifically selected as it efficiently handles data distributions with different scales and is particularly effective for models sensitive to feature magnitudes, such as support vector machines and gradient boosting algorithms. Column names were standardized by removing unwanted characters and replacing problematic symbols (e.g., '<' replaced by '_less') to guarantee compatibility with the machine learning libraries used, notably XGBoost.

Iterative Refinement and Validation

Throughout the modelling process, initial model training highlighted additional data cleaning needs, prompting further refinements. For example, early model runs revealed issues with column names, leading to us standardising the column names, as mentioned above. This iterative cycle of training, analysis, and refinement significantly enhanced dataset quality and model robustness.

After completing the data cleaning processes, we conducted exploratory data analyses and descriptive statistical checks to ensure that the preprocessing steps had not introduced unintended biases or errors. This validation was crucial in maintaining the integrity and reliability of the dataset for subsequent modelling.

Limitations and Future Considerations

Despite careful data cleaning and preprocessing, our approach presented some limitations. Aggressive removal of certain features, although justified, might inadvertently discard potentially valuable insights. Future projects could explore advanced imputation techniques, dimensionality reduction methods such as Principal Component Analysis (PCA), or alternative feature-selection frameworks to balance data integrity and complexity reduction more effectively.

Finally, we acknowledge that bootstrapping the test set results could have provided additional insight into the sampling variability of our performance metrics. Resampling the holdout season would have delivered empirical confidence intervals around ROI and accuracy, helping to judge whether observed gains are likely to generalise to future seasons. Implementing a full bootstrap workflow was, however, beyond the scope for this project, so we report point estimates only and flag this as an avenue for later work.

Machine Learning Models

Leveraging eight algorithms plus an ensemble voter reflects our commitment to methodological pluralism by letting multiple imperfect lenses triangulate market behaviour.

Logistic Regression

Logistic regression is a widely used classification algorithm, particularly effective in predicting binary or multinomial outcomes. It functions by estimating probabilities through the logistic or

sigmoid function, converting any input value into a probability between 0 and 1. This approach assumes a linear relationship between input variables and the log-odds of the predicted outcome (James et al., 2021). The model parameters are typically estimated using maximum likelihood estimation, aiming to find coefficients that maximize the likelihood of observed outcomes.

Unlike tree-based algorithms such as Random Forest or XGBoost, logistic regression offers greater interpretability through direct feature importance via its coefficients. It explicitly illustrates the influence of each predictor on the outcome probability, providing clarity that complex models often lack (Hastie, Tibshirani & Friedman, 2009). This interpretability is particularly beneficial in domains where understanding underlying factors is as crucial as predictive accuracy, such as sports analytics.

For our research into Premier League (PL) outcomes and betting strategies, logistic regression stands out due to its robust performance in generating well-calibrated probabilities (Wijk, 2021). Since our goal is to identify mismatches between predicted probabilities and bookmaker odds, the logistic regression model's inherent probability estimates can directly inform betting decisions. For instance, if our logistic model estimates a higher probability of a home win than bookmakers imply, it signals a valuable betting opportunity. Van Wijk (2021) demonstrated that logistic regression not only performs robustly in predicting football match results but can also consistently identify profitable betting scenarios by exploiting bookmaker inaccuracies. Given its proven effectiveness and interpretability, logistic regression serves as a solid baseline and valuable analytical tool for your research, facilitating clearer strategies for betting profitably.

Random Forest

Random Forest is a versatile and powerful ensemble method used in both regression and classification tasks. It builds multiple decision trees by repeatedly sampling subsets of training data (bootstrap samples) and features, resulting in a robust model capable of capturing complex interactions and reducing overfitting through averaging predictions from individual trees (Breiman, 2001). Each tree independently learns patterns from data, and their predictions are

combined through majority voting (classification) or averaging (regression), improving predictive stability and accuracy.

Compared to simpler linear models such as logistic regression, Random Forest can automatically detect nonlinear relationships and interactions without needing explicit feature transformations. Its ability to model intricate, high-dimensional data makes it particularly valuable when predicting sports outcomes, as the interactions among factors like team form, and past performance are rarely linear or straightforward (Breiman, 2001).

In our research on Premier League predictions, Random Forest can uncover hidden patterns and identify key predictors that simpler methods might overlook. While Van Wijk (2021) found Random Forest slightly less profitable than logistic regression, the model excelled in highlighting important predictive features, such as recent performance metrics. These insights are valuable when refining betting strategies, as Random Forest's detailed feature-importance metrics provide guidance for selecting or discarding factors. Thus, Random Forest offers both predictive power and analytical depth, supporting nuanced betting strategies by indicating critical aspects that influence match outcomes.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple yet intuitive classification algorithm that assigns class labels based on the most common outcome among the nearest points in the feature space. Unlike parametric models such as logistic regression or tree-based algorithms, KNN does not create explicit predictive functions during training. Instead, it stores all training data points, delaying computation until a prediction is required (Cover & Hart, 1967). This approach classifies a new data point based solely on its proximity to previously observed cases.

KNN's simplicity and instance-based nature distinguish it significantly from other more sophisticated algorithms. It relies heavily on the choice of the parameter 'k' (number of neighbors considered), distance metrics, and feature scaling. The algorithm performs well when data points

of similar classes naturally cluster together, making it valuable in scenarios where historical outcomes consistently repeat under similar conditions (James et al., 2021).

For our Premier League predictions and betting research, KNN may serve as a useful baseline model, particularly for its transparency and interpretability. Although KNN is often overshadowed by more powerful methods, it still effectively captures local similarities. For example, matches involving teams with historically comparable form could be identified accurately by KNN, potentially revealing betting opportunities missed by global models. Nonetheless, due to limitations when handling numerous irrelevant or noisy features, careful feature selection and normalization are essential for KNN to deliver useful predictive insights in football analytics (James et al., 2021).

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an advanced tree-based ensemble model recognized for its superior performance and efficiency. It sequentially constructs small decision trees, each correcting errors made by previous ones, through gradient descent optimization of a loss function (Chen & Guestrin, 2016). XGBoost's strength lies in its capability to handle complex interactions and nonlinear patterns, common characteristics of sports-related data, making it popular for predictions involving structured and diverse datasets.

Unlike Random Forest, which independently averages predictions, XGBoost strategically corrects misclassifications, often resulting in greater accuracy. Its robust regularization options, such as pruning and shrinkage, mitigate overfitting, and its computational efficiency allows effective handling of large-scale data, important considerations for extensive historical football data (Bentéjac, Csörgő & Martínez-Muñoz, 2021).

In our study aiming to gain an advantage on the bookmakers by predicting Premier League outcomes, XGBoost's accuracy in handling intricate, multivariate interactions among match statistics and conditions can significantly enhance betting decisions. Given our objective to

identify profitable odds discrepancies, XGBoost's refined probability predictions offer a substantial advantage in spotting betting value through intricate pattern recognition.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification algorithm that finds an optimal decision boundary (hyperplane) between classes by maximizing the margin between them (Cortes & Vapnik, 1995). It employs kernel functions, allowing nonlinear data transformations, thus enabling it to separate complexly interwoven classes in high-dimensional spaces.

Distinctively, SVM does not inherently provide probability estimates, differing from logistic regression and ensemble models. However, techniques such as Platt scaling can convert its outputs into approximate probabilities, though these methods require additional steps and careful calibration (James et al., 2021).

For predicting football outcomes and beating bookmakers, SVM's primary advantage lies in its ability to handle intricate class boundaries effectively. Nevertheless, given its computational intensity and complexity in tuning kernels and hyperparameters, it may be less practical compared to logistic regression or tree-based models, particularly if clear probabilistic interpretations and swift recalibrations (such as weekly match updates) are essential to our betting strategy.

Naive Bayes

Naive Bayes is a simple probabilistic classifier based on Bayes' theorem, operating under the assumption that each feature independently contributes to the probability of an outcome. Despite this simplifying assumption, which rarely holds true perfectly, Naive Bayes often performs surprisingly well, particularly with small to medium-sized datasets or in high-dimensional settings (James et al., 2021). The model calculates the likelihood of each class based on feature distributions and combines this with prior class probabilities to produce posterior probabilities for prediction.

What differentiates Naive Bayes from models like logistic regression or ensemble methods is its simplicity and speed, as it does not require extensive parameter tuning or complex computations. However, its main limitation is the assumption of independence between predictors, which can reduce accuracy when features are highly correlated (James et al., 2021).

In the context of our research on Premier League betting outcomes, Naive Bayes might serve as a useful benchmark or complementary method. While Van Wijk (2021) did not directly test Naive Bayes, literature cited in the thesis indicated moderate success in football predictions. Its advantage lies in generating clear and interpretable probability estimates, enabling quick identification of discrepancies against bookmaker odds. Nevertheless, its assumptions might oversimplify complex interactions present in football data (Wijk, 2021).

Light Gradient Boosting Machine (LightGBM)

LightGBM (Light Gradient Boosting Machine) is a fast and highly efficient gradient boosting framework that builds decision trees to achieve accurate predictive performance. It employs histogram-based algorithms and leaf-wise tree growth, resulting in faster training speed and lower memory consumption compared to traditional gradient boosting methods (Ke et al., 2017). Additionally, LightGBM utilizes Gradient-Based One-Side Sampling (GOSS) to prioritize data points where errors are largest, enhancing its predictive accuracy with fewer resources.

Compared to XGBoost or Random Forest, LightGBM typically delivers similar or better accuracy in less computational time, especially suitable for large or high-dimensional datasets. Its capability to handle numerous features efficiently makes it particularly attractive when extensive match statistics, or historical odds are considered (Bentéjac, Csörgő & Martínez-Muñoz, 2021).

Given LightGBM's proven effectiveness in recent sports analytics literature, it holds significant promise for improving betting accuracy and profitability. For our study, LightGBM's efficient handling of data and high predictive accuracy might lead to quicker identification of profitable

betting opportunities, particularly when rapid model updates (e.g., weekly matches) are essential to consistently outperform bookmakers (Bentéjac, Csörgő & Martínez-Muñoz, 2021).

CatBoost

CatBoost is a specialized gradient boosting algorithm designed explicitly to handle categorical variables without requiring extensive feature engineering or encoding. Developed by Yandex, CatBoost applies an advanced technique called ordered target encoding, carefully managing target leakage, and significantly improving accuracy when numerous categorical predictors are involved (Prokhorenkova et al., 2018).

Its unique ability to directly manage categorical data differentiates it from methods like XGBoost or LightGBM, which typically require manual categorical encoding, thus increasing complexity or dimensionality. CatBoost's built-in regularization techniques and ordered boosting method make it highly resistant to overfitting, ensuring reliable probability estimates (Bentéjac, Csörgő & Martínez-Muñoz, 2021).

For predicting Premier League outcomes and informing profitable betting strategies, CatBoost's capability to naturally incorporate categorical variables, such as team identities, referees, stadiums, or matchups, can provide distinct predictive advantages. Literature strongly suggests CatBoost’s superior performance in sports outcome prediction tasks involving categorical data. Therefore, integrating CatBoost could significantly enhance our research, efficiently uncovering subtle, category-specific betting opportunities overlooked by bookmakers (Wijk, 2021).

Voting

A Voting classifier is an ensemble method that combines predictions from multiple diverse models, such as logistic regression, decision trees, and support vector machines, to achieve improved accuracy and robustness. The ensemble prediction is determined either through hard voting (majority class votes) or soft voting (averaging class probabilities across models) (Polikar, 2006). Initially, a hard voting classifier was tested, but ultimately a soft voting approach was adopted in our method due to its superior predictive effectiveness.

Unlike methods such as Random Forest or XGBoost that aggregate similar models, a Voting classifier merges distinct algorithms, each capturing unique patterns in data. This diversity enables the ensemble to compensate for individual model weaknesses, potentially increasing overall predictive accuracy and reliability. It is particularly effective when models complement each other, ensuring errors are less correlated and accuracy enhanced (James et al., 2021).

In the context of Premier League outcome predictions, a Voting classifier could integrate Van Wijk's best-performing logistic regression model with other algorithms, potentially further improving predictive outcomes and betting profitability. Van Wijk (2021) did not explicitly test Voting classifiers; however, previous studies highlighted in the thesis acknowledge ensemble approaches' value in boosting predictive performance for sports analytics. By carefully selecting and combining complementary models, our research could use Voting classifiers to stabilize predictions, reduce variance, and improve decision-making quality when identifying profitable betting opportunities (Wijk, 2021).

Model Tuning and Refinement

Initial Approach to Hyperparameter Tuning

Initially, we adopted a conventional hyperparameter tuning approach, employing Bayesian optimization due to computational constraints. Bayesian optimization refers to an iterative method for optimizing objective functions (in this case, model performance metrics) using a surrogate model, typically a Gaussian process or a tree-based surrogate (Bischl et al., 2023). We used a Gaussian process. Rather than exhaustively searching or randomly sampling hyperparameters, Bayesian optimization strategically explores the hyperparameter space by predicting the performance of different configurations based on prior evaluations, then selecting the most promising candidates for subsequent trials (Bischl et al., 2023). This targeted approach significantly reduces the number of evaluations needed, making it especially suitable when computational resources are limited. It was chosen for its efficiency and effectiveness in optimizing hyperparameters within a limited number of trials compared to methods like grid

search or random search. Our primary objective at this stage was to tune each model to achieve the highest possible accuracy on predicting match outcomes.

Transition to ROI-focused Hyperparameter Tuning

Upon seeing the lack of improvement to the model's return on investment when using a flat betting strategy post tuning, we recognized that our initial focus on maximizing predictive accuracy did not align with the project's primary objective, maximizing return on investment (ROI). Achieving high accuracy does not necessarily translate to greater profitability, as correctly predicting outcomes with lower odds yields less financial return than correctly predicting fewer, less probable outcomes with higher odds.

To address this misalignment, we developed a custom hyperparameter tuning strategy, explicitly optimizing hyperparameters for maximum ROI rather than accuracy alone. Re-targeting optimisation to ROI therefore makes our axiological focus on actionable profit explicit inside the tuning routine. Initially, we used the test dataset for tuning, which constituted data leakage and could artificially inflate our models' perceived profitability. Realizing this, we corrected our methodology by allocating a validation set derived from approximately 10% of our training dataset (roughly half a season of games). This validation set allowed us to fine-tune our hyperparameters ethically and effectively, without introducing bias from the test data, nor risking overfitting to the training data. We acknowledge that using a small validation set for hyperparameter tuning introduces a risk of model's overfitting to said set, however this can be identified through poor results on the test set, which was not the case, and therefore we can confidently say there was limited overfitting to the validation set.

The tuning continued with Bayesian optimization with a Gaussian process, as computational resources remained limited. Each model, Random Forest, XGBoost, LightGBM, CatBoost, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN), underwent optimization through ROI-based evaluation through a flat betting strategy. The optimization process iteratively adjusted hyperparameters to identify configurations delivering the highest potential financial returns, measured by the ROI calculated from simulated betting outcomes based on predicted probabilities and bookmaker odds.

Additionally, the Naive Bayes model required no hyperparameter tuning due to its intrinsic simplicity and lack of configurable parameters, providing a baseline comparison for more complex models.

Possible Additional Considerations

Future research could explore extending the validation set further to cover different football seasons or incorporate cross-validation techniques to enhance generalizability. Another potential improvement could involve dynamically adjusting the optimization criteria to balance ROI maximization with model robustness, ensuring profitable outcomes are not overly dependent on a few high odds matches. Finally, we could also have optimized the models' parameters for each betting strategy, however, we deemed it out of scope for this project, due to the time-consuming nature of the task with limited computational resources.

Betting Strategies

In sports betting, predictive accuracy alone does not guarantee profitability. This section will be highlighting which betting strategies will be explored to yield higher returns when used in conjunction with the machine learning models. Return on Investment (ROI) is the primary metric we use to measure success, as it directly quantifies profitability relative to the amount wagered, effectively capturing both prediction accuracy and betting efficiency. Using ROI allows us to clearly determine whether machine learning predictions can consistently gain an advantage on bookmakers and translate predictive insights into tangible financial outcomes.

Flat Betting

Flat betting is a straightforward strategy involving consistently betting the same amount of money on each wager, irrespective of the perceived probability or potential return from the match outcome. Under this approach, we utilize a fixed unit stake, for example, $1, and wager exactly this amount on every single game as predicted by each machine learning (ML) model.

Unlike progressive or proportional strategies, flat betting does not adjust stakes based on previous outcomes, confidence levels, or odds offered. Its simplicity is beneficial, allowing us to directly assess the performance of prediction models without the confounding factor of varying stakes (Wijk, 2021).

In practical application, flat betting treats every predicted match equally. Whether the match features a clear favourite or a highly uncertain outcome, the same predetermined stake is placed on the result predicted by the ML model. This uniformity simplifies performance assessments such that, if a prediction is correct, returns are directly proportional to the bookmaker's odds. If incorrect, the stake is simply lost. As a result, flat betting is frequently used as a baseline strategy in academic and practical evaluations, since any differences in profitability can be attributed entirely to the predictive quality of the model rather than to variable stakes (Rue & Salvesen, 2000).

Flat betting is highly relevant to research into maximizing return on investment (ROI) using ML predictions for Premier League matches. When applying ML models, we aim to measure whether the predictions genuinely outperform the market odds. Flat betting facilitates this comparison by standardizing stake size and isolating predictive accuracy from bankroll management effects. Although more sophisticated strategies may offer theoretical improvements, flat betting's simplicity provides clarity and consistency in evaluation, making it an ideal starting point (Wijk, 2021). While other more nuanced money management techniques may yield a higher ROI, flat betting is the baseline strategy used to compare any utilized betting strategy.

Threshold Betting

Threshold betting introduces selectivity to wagering, placing bets only when the predicted probability for a particular match outcome surpasses a predetermined confidence threshold. This approach focuses specifically on predictions that the ML models identify as highly confident. For example, we might decide to wager only when the model's predicted probability exceeds 60% or another set benchmark (Forrest & Simmons, 2001). Adjusting this threshold doubles as a live calibration test, echoing the post positivist demand for continual attempts at refutation.

Operationally, we would evaluate each prediction from the ML model against this predetermined confidence threshold. Bets are only placed on matches surpassing this threshold, while all other opportunities are disregarded, regardless of perceived potential odds value. Consequently, the total volume of bets decreases significantly, with each wager now representing a higher-confidence prediction. This selective strategy aims to increase the success rate of placed wagers and thereby potentially enhance overall ROI. In this way, the threshold method acts as a systematic filter, restricting betting activity to the scenarios the ML models are most confident about predicting accurately (Rue & Salvesen, 2000).

Threshold betting is critically relevant to research into ML-driven betting strategies because it tests the quality and reliability of the ML models' confidence assessments. By only engaging in betting when predictive confidence is high, we can directly analyse how model calibration impacts profitability. If increasing the threshold improves ROI, it signals robust model calibration. However, too high a threshold can reduce betting opportunities excessively, thereby limiting profit potential. Hence, threshold betting serves as a practical, strategic framework for evaluating the trade-off between prediction confidence and betting frequency (Forrest & Simmons, 2001), meaning the most important consideration here is determining where to place the confidence threshold. It must not be so high as to limit potential profits, yet not so low that too many lower confidence bets are placed, hypothetically increasing the number of lost wagers.

Value Betting

Value betting strategically targets opportunities where the odds provided by bookmakers exceed the "fair" odds implied by an ML model's predicted probabilities. This method explicitly seeks out discrepancies between the bookmaker's odds and the probabilities calculated by the ML models. In essence, value betting identifies bets where bookmakers have potentially mispriced the odds, offering returns greater than justified by the model's predicted outcomes (Constantinou & Fenton, 2013).

In practice, value betting involves comparing the ML-generated probability directly against bookmaker odds. For example, if the model estimates a 40% chance of a team winning (equivalent to fair odds of 2.50), but the bookmaker offers odds of 3.00 (implying a probability of around 33.3%), the difference constitutes betting value. In these cases, a wager is placed because the expected value (probability multiplied by payout) is greater than the cost of placing the bet, theoretically providing positive long-term returns (Vlastakis, Dotsis & Markellos, 2006). Naturally, this means that each ML model must determine its own fair odds, compare them to the odds of the bookmaker and only if there is a larger expected value, place the wager. This also means considerations into changing odds arise. Odds can change prior to games commencing, meaning when to bet and when the models consider each bookmaker's odds become a factor.

Value betting is highly relevant to our research in employing ML predictions to forecast Premier League outcomes because it directly exploits model insights to identify market inefficiencies. This strategy allows us as bettors to rigorously assess whether the ML models' probability predictions truly offer an advantage over bookmaker-generated odds. As the central premise of value betting revolves around recognizing when bookmakers underestimate or overestimate probabilities, it effectively tests the quality of the ML models' predictive accuracy and calibration. By strategically identifying and leveraging such discrepancies, value betting can significantly enhance ROI, providing clear empirical evidence of predictive value (Constantinou & Fenton, 2013).

Kelly Criterion Betting

The Kelly Criterion, originally based on work published by Kelly in 1956, is a betting strategy that aims to maximize long-term capital growth by adjusting the stake based on any betting advantage proportional to the size of the advantage. Unlike flat betting, stakes vary dynamically, depending on how strongly a model's predicted probability surpasses the odds implied by bookmakers. A key theoretical principle behind the Kelly Criterion is optimizing growth rate by precisely calculating the amount wagered to reflect the estimated betting edge (Thorp, 2006). Practically, the Kelly Criterion formula is straightforward:

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

where f is the fraction of the current bankroll to wager, p is the model's estimated confidence of success, and is the net odds offered by bookmakers (Kelly Jr, 2011). If the resulting fraction is positive, the bettor places that percentage of their bankroll on the wager, otherwise, no bet is made. As the edge increases, so does the wagered amount. However, this approach depends heavily on the accuracy and calibration of the model's probability estimates. Overly optimistic predictions may lead to unnecessarily large wagers, while overly conservative estimates may miss profitable opportunities. In practice we used a fractional-Kelly approach, wagering only 0.3 of the model's bankroll (30 % of the full-Kelly stake) on each qualifying bet to tame variance and drawdowns. An important note is the net odds offered by bookmakers are used in this Kelly Criterion strategy in terms of the probability of the event occurring. This assumes bookmakers use completely fair odds which can be directly translated to probabilities. However, bookmakers incorporate overrounds in their odds, to ensure they have the statistical advantage (Newall, 2015), similar to the house edge used by casinos.

In the context of using ML predictions for Premier League outcomes, the Kelly Criterion is particularly relevant because it leverages the quality and precision of probability predictions produced by these models. If an ML model can accurately predict match outcomes better than bookmakers, the Kelly strategy theoretically provides an optimal method for stake allocation. However, practical applications need caution, as even minor inaccuracies in probability estimation can significantly influence outcomes. Thus, it serves as an important methodological consideration in assessing whether ML models yield robustly profitable betting outcomes in sports betting research (Thorp, 2006).

Ethical considerations

Gambling carries a real risk of addiction and financial loss. The algorithms described here are shared for research and education only. Anyone who applies them should stake only money that he or she has deliberately accepted may fall to zero, set strict bankroll limits, and use

available self-exclusion or cooling-off tools. All activity must comply with the gambling laws and regulations in the relevant jurisdiction. Responsible use also means tracking outcomes, watching for signs of harm, and pausing or ending wagering if negative effects emerge.

Results

This section presents the results obtained by applying the various machine learning models and betting strategies as described in the Methodology section. Specifically, we focus on two primary performance metrics, prediction accuracy and ROI. Prediction accuracy evaluates the effectiveness of each model in correctly forecasting match outcomes, while ROI assesses the profitability achieved using different betting and stake-management approaches. The following subsections will provide detailed comparisons of model performances, identify the most profitable betting strategies, and discuss notable patterns or anomalies revealed through the analysis of Premier League data from the selected five-season period.

Classification Performance

Model	Accuracy	Precision	Recall	F1	AUC	Brier
Random Forest	0.58	0.38	0.49	0.43	**0.71**	**0.55**
Logistic Regression	0.57	0.54	0.47	0.42	0.69	**0.55**
XGBoost	0.56	0.47	0.47	0.44	0.70	0.60
SVM	0.56	0.37	0.47	0.41	0.66	0.59
KNN	0.55	0.44	0.46	0.43	0.67	0.57
Naive Bayes	0.51	0.49	0.49	**0.48**	0.69	0.94
LightGBM	0.54	0.44	0.47	0.44	0.70	0.60
CatBoost	**0.59**	**0.61**	**0.50**	0.45	0.70	**0.55**
Dummy Classifier	0.47	0.16	0.33	0.21	0.50	1.00
Voting	**0.59**	0.46	**0.51**	0.46	**0.71**	**0.54**

***Table 4**: Metrics per model evaluated.*

Table 4 summarizes the performance metrics for each model evaluated. *Accuracy* measures the proportion of correctly predicted outcomes, *Precision* shows the proportion of positive identifications that were correct, and *Recall* measures the proportion of actual positives correctly identified by the model. The *F1-score* is the harmonic mean of precision and recall, combining an average of both metrics into a single score. The *Area Under the Curve* (AUC) evaluates a model's ability to distinguish between classes across different classification thresholds, with a higher AUC indicating better overall performance. The *Brier score* assesses the accuracy of predicted probabilities, where lower scores indicate more accurate probabilistic predictions. Brier scores range between 0 and 1, with a score of 0 representing perfect accuracy and a score of 1 representing complete inaccuracy. Precision, recall, F1 and AUC are all macro-averaged, meaning the metrics were computed for each class separately, and then averaged equally across classes so each one contributes the same weight to the overall score.

Among all models assessed, CatBoost achieved the highest accuracy of 0.59 and precision of 0.61, indicating it correctly identified match outcomes most effectively and had the lowest false-positive rate. Conversely, the Dummy Classifier, which serves as our baseline, performed poorest with an accuracy of only 0.47, precision of 0.16, and significantly weaker performance across all metrics. Interestingly, despite its high accuracy, CatBoost's F1-score of 0.45 reveals a notable imbalance between precision and recall, suggesting the model may be biased towards certain classes.

The Random Forest model showed competitive performance with an accuracy of 0.58 and the highest AUC of 0.71. The strong AUC indicates Random Forest’s effectiveness in distinguishing between outcomes across the full spectrum of classification thresholds. The Logistic Regression and Naive Bayes models performed similarly with slightly lower accuracy (0.57 and 0.51 respectively) and an identical AUC of 0.69. Notably, Naive Bayes displayed a particularly high Brier score (0.94), highlighting its weaker probabilistic prediction accuracy despite decent classification performance. Finally, as the dummy model always assigns 100 percent probability to class 0, its squared error is 0 for the half of test samples that truly are class 0 and 2 for the rest, so the average Brier score is 1.

XGBoost, LightGBM, and KNN all showed similar middle of the range performances, with accuracy scores ranging between 0.54 and 0.56, and AUC scores between 0.67-0.69. The moderate AUC values suggest potential for further improvements, through more extensive hyperparameter optimization or refined feature selection. While we performed Bayesian optimization due to computational limitations, a more exhaustive grid search could yield additional performance gains The Support Vector Machine (SVM) model demonstrated relatively lower precision (0.37) but maintained a similar accuracy and recall to other mid-range models, implying it may be more prone to class imbalance or feature noise issues.

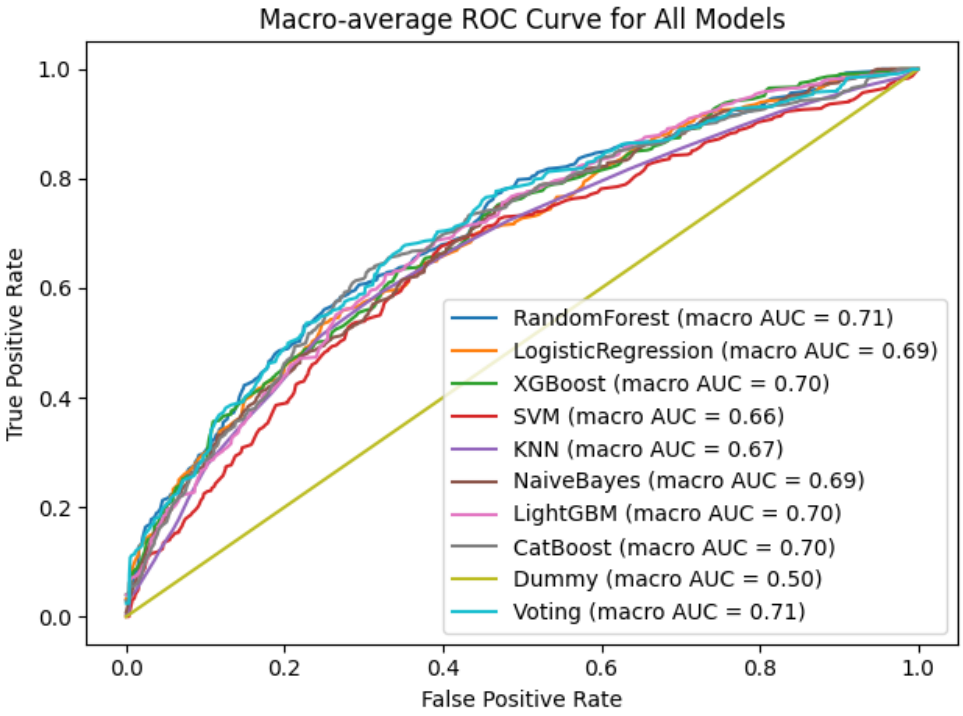


Figure 4: *Macro-average ROC curve for all models evaluated.*

Figure 4 illustrates the relative predictive capabilities of each model via their ROC curves. The ROC curve (Receiver Operating Characteristic curve) shows how well each model can distinguish between different outcomes by comparing its correct predictions against its false predictions at various points. Since football games can result in three outcomes for a given team (Win, Loss, or Draw), the baseline random accuracy would theoretically be 33.3% in a three-class problem, representing a purely random prediction without any informative power.

However, the diagonal dummy line in the ROC plot specifically represents the performance of a completely uninformative classifier, which always yields an AUC of 0.5, regardless of how many classes we have. Even though there are three outcomes, the AUC baseline remains at 0.5 because it measures how well the model ranks outcomes against each other (one-vs-rest comparisons), not just how often it guesses exactly right. Random guessing will always correctly rank outcomes half the time, regardless of the total number of outcomes. Therefore, models displaying ROC curves significantly above this diagonal line, such as Random Forest, demonstrate a clear advantage in prediction capability. Models closer to this diagonal line illustrate lower predictive power.

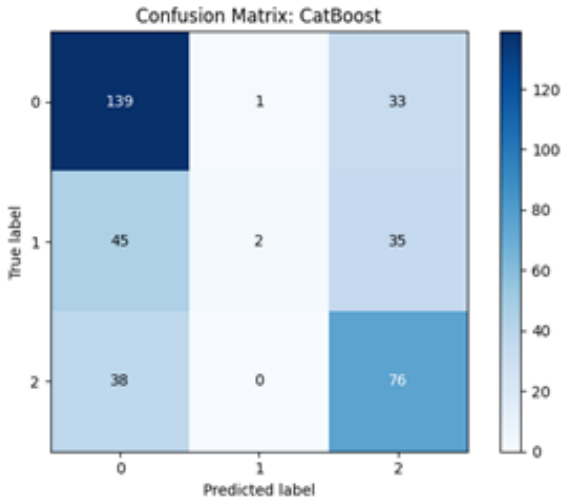


***Figure 5**: Confusion Matrix, CatBoost*

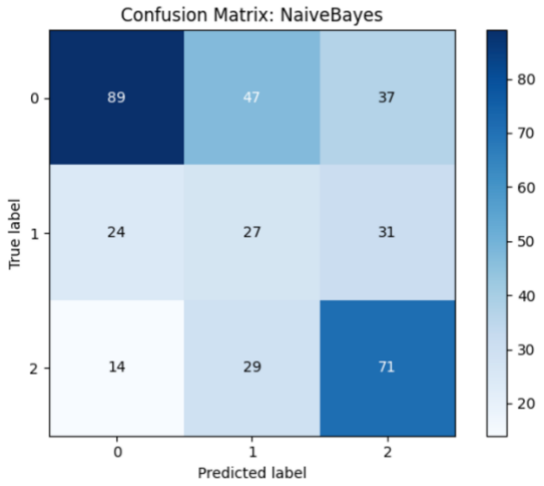


***Figure 6**: Confusion Matrix, Naive Bayes*

The confusion matrices for the CatBoost (Figure 5) and Naive Bayes (Figure 6) models further show the predictive strengths and weaknesses. CatBoost, the highest performing model, managed to correctly classify 139 home wins and 76 away wins, showing impressive performance in distinguishing these two outcomes. However, it misclassified 45 draws as home wins, and 35 as away wins, reflecting challenges in differentiating draws from wins. In contrast, Naive Bayes, the lowest performing model, misclassified more outcomes, only correctly identifying 89 home wins while incorrectly classifying 47 home wins as draws, showing substantial difficulty in distinguishing between these classes. The Naive Bayes model was better at predicting draws (27 correct draw predictions) compared to CatBoost (2 correct draw predictions). However, out of 82

draws, correctly predicting 27 is still not optimal. In contrast, the CatBoost model excels at predicting home wins, correctly predicting 139 out of 173 home wins, a success rate of 80.35%.

The confusion matrices for the other models (Appendix 1), such as Logistic Regression and SVM models also illustrate the predictive strengths and weaknesses. Logistic Regression demonstrates relatively strong performance, correctly identifying home wins effectively but struggling with accurately predicting draws and away wins. Conversely, the SVM model exhibits substantial weaknesses, consistently misclassifying outcomes, particularly failing to correctly predict draws altogether.

Feature Importance

Across our ensemble of machine-learning models, two clear patterns emerged (see Appendix II for feature importance table). First, pre-match corner-odds, in particular Pinnacle's home/draw/away corner markets and market-maximum corner odds, consistently ranked as the top three most important features in all tree-based learners (Random Forest, XGBoost, LightGBM, CatBoost) and in the soft-voting ensemble. Second, engineered "recent form" metrics, such as home and away goal-conversion rates over the past one to five matches and foul-to-card ratios over the past three to ten matches, appeared in the top five for each non-linear model. Linear models (Logistic Regression) instead emphasized recent disciplinary stats (fouls and red cards in the last match) and Asian-handicap pricing, while instance-based learners (SVM, KNN) focused respectively on detailed form statistics and over/under goal markets. Naive Bayes, by contrast, learned directly from categorical team identifiers and interestingly flagged previous matches' red cards as influential for outcome prediction.

Betting Simulation Performance

In this subsection, we show how each of the four betting strategies performed when applied to the predictions made by each machine learning model. The evaluation is conducted through simulated betting over the test period, evaluating how each model and strategy combination affects the hypothetical betting balance. Equity curves will be used to visually depict how the betting balances evolved over time, clearly showing periods of gains and losses, and thus

providing insight into the consistency and volatility of each model's performance under each betting strategy. Each simulation began with a starting balance of $100 (100 units), chosen both as a plausible wallet size for typical retail bettors and to allow straightforward comparative analyses between models and strategies.

Strategy	Model	Bets Placed	Win %	Total Staked	Net Profit	ROI %	Sharpe	Max Drawdown
Flat	Random Forest	369	57.72	369.00	1.19	1.19	0.21	0.12
Flat	Logistic Regression	369	56.91	369.00	-1.40	-1.40	0.07	0.14
Flat	XGBoost	369	55.56	369.00	-4.38	-4.38	-0.1	0.20
Flat	SVM	369	56.37	369.00	17.11	17.11	0.9	0.18
Flat	KNN	369	55.28	369.00	-12.27	-12.27	-0.52	0.17
Flat	Naive Bayes	369	50.68	369.00	-1.08	-1.08	0.11	0.25
Flat	LightGBM	369	54.47	369.00	6.34	6.34	0.46	0.13
Flat	CatBoost	369	58.81	369.00	13.72	13.72	0.84	0.11
Flat	Voting	369	58.81	369.00	19.19	19.19	1.24	0.10
Threshold	Random Forest	132	75.00	132.00	10.93	10.93	1.65	0.04
Threshold	Logistic Regression	134	74.63	134.00	9.99	9.99	1.36	0.04
Threshold	XGBoost	259	63.32	259.00	12.10	12.10	1.03	0.11
Threshold	SVM	29	68.97	29.00	-0.28	-0.28	-0.06	0.03
Threshold	KNN	159	70.44	159.00	11.18	11.18	1.35	0.06
Threshold	Naive Bayes	364	50.82	364.00	-0.71	-0.71	0.13	0.25
Threshold	LightGBM	234	59.40	234.00	-1.06	-1.06	-0.01	0.12
Threshold	CatBoost	132	69.70	132.00	0.41	0.41	0.09	0.06
Threshold	Voting	168	74.40	168.00	15.91	15.91	2.06	0.05
Value	Random Forest	161	54.04	161.00	4.87	4.87	0.44	0.11
Value	Logistic Regression	181	55.25	181.00	1.51	1.51	0.17	0.09
Value	XGBoost	324	55.86	324.00	4.19	4.19	0.4	0.15
Value	SVM	131	36.64	131.00	3.50	3.50	0.28	0.19
Value	KNN	220	46.36	220.00	-20.22	-20.22	-1.14	0.22
Value	Naive Bayes	368	50.54	368.00	-1.81	-1.81	0.08	0.25
Value	LightGBM	307	53.42	307.00	11.84	11.84	0.7	0.13
Value	CatBoost	167	53.29	167.00	9.85	9.85	0.84	0.11
Value	Voting	222	56.31	222.00	21.32	21.32	1.71	0.08
Kelly	Random Forest	161	54.04	747.49	61.28	61.28	1.19	0.29
Kelly	Logistic Regression	181	55.25	945.39	0.13	0.13	0.45	0.60
Kelly	XGBoost	324	55.86	6268.71	74.20	74.20	1.37	0.91
Kelly	SVM	131	36.64	1025.62	4.12	4.12	0.59	0.77
Kelly	KNN	220	46.36	916.10	-67.24	-67.24	-0.34	0.77

Kelly	Naive Bayes	79	49.37	669.71	-100.00	-100.00	-0.24	1.00
Kelly	LightGBM	307	53.42	1140.28	-80.69	-80.69	0.64	0.96
Kelly	CatBoost	167	53.29	678.33	83.15	83.15	1.38	0.34
Kelly	Voting	222	56.31	1426.68	133.63	133.63	1.52	0.53

Table 5: *Summary of Strategy and Models*

The results highlight several key findings about model performances across different betting strategies (Table 5). Unsurprisingly, the Voting model consistently outperformed other models, achieving substantial returns particularly under the Kelly Criterion strategy, where it reached an ROI of 133.63%. Similarly, CatBoost and XGBoost also demonstrated strong capabilities, with CatBoost reaching an ROI of 83.15% under Kelly conditions, and XGBoost maintaining profitability across multiple strategies, particularly under Threshold Betting (12.10% ROI) and Kelly Criterion (74.20% ROI). On the other hand, models like KNN, Naive Bayes, and LightGBM frequently struggled. KNN exhibited significant challenges, particularly under Value (-20.22% ROI) and Kelly (-67.24% ROI) strategies, indicating a substantial misalignment between its probability predictions and bookmaker odds. Naive Bayes similarly faced major difficulties, losing its entire bankroll under Kelly Criterion conditions (ROI of -100%), and regularly producing negative results across other strategies.

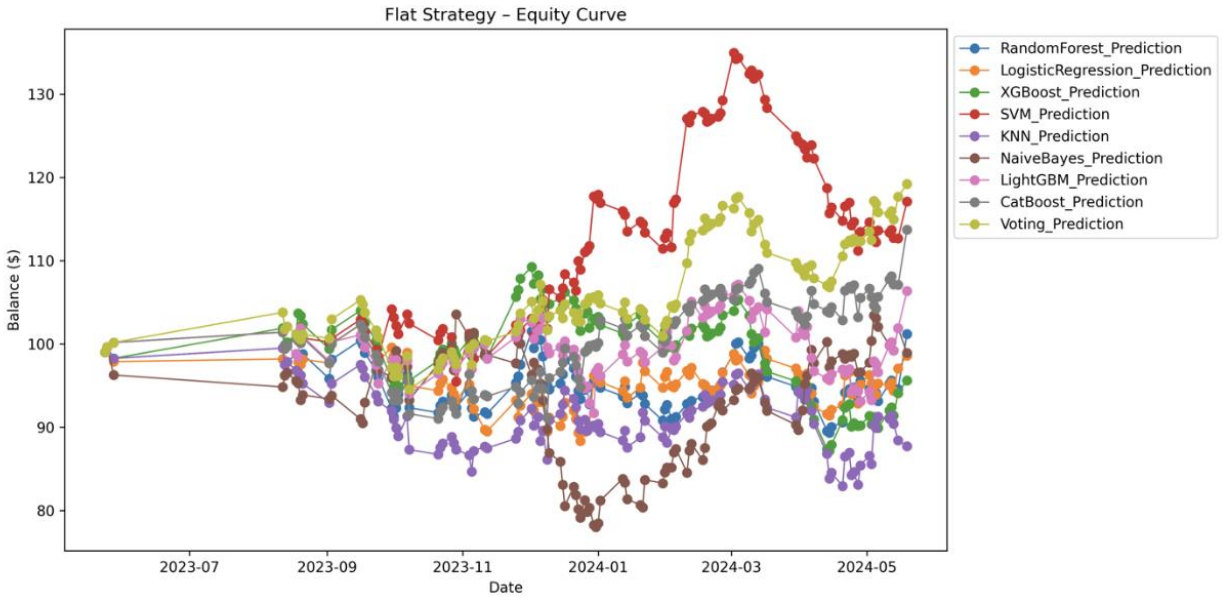


Figure 7: *Flat Betting Equity Curve*

The Flat Betting strategy provided a clear baseline for performance evaluation, as it involves betting an equal, fixed amount on every predicted match outcome, regardless of the confidence level or perceived value. When reviewing the flat betting equity curve (Figure 7), the model that stands out is SVM, at certain points soaring above a $130 balance. The models that performed best under flat betting conditions were Voting ($119.19 closing balance), SVM ($117.11 closing balance) and CatBoost ($113.72 closing balance). Conversely, the model which significantly underperformed under flat betting conditions was KNN, finishing with a $87.73 closing balance, representing a -12.27% ROI over the period. XGBoost, Logistic Regression, and Naive Bayes also yielded negative ROIs however, neither of them exceeding a loss of greater than 5%, with -4.38%, -1.4%, and -1.08% respectively.

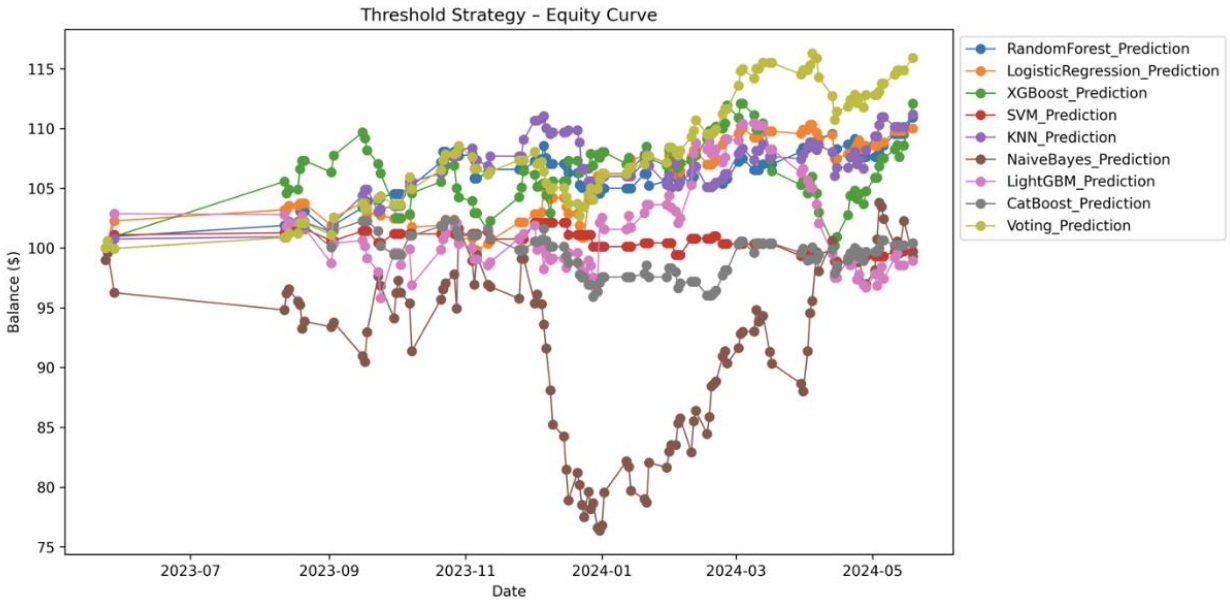


Figure 8: *Threshold Betting Equity Curve*

Threshold Betting involved placing bets only when the model's predicted probability exceeded a certain confidence threshold (60% confidence). This selective approach reduced betting frequency, aiming to focus only on highly confident prediction. This meant that models were betting on a lower quantity of games, with the bets placed ranging from 29 (SVM) to 364 (Naive Bayes), out of a total 369 potential bets. The threshold betting equity curve (figure 8), shows the Naive Bayes model having a significant downturn (25% Max Drawdown), suggesting challenges in reliable confidence estimation, leading to unprofitable outcomes even when predictions were

deemed highly confident. Naive Bayes did, however, despite this large downturn, and a max drawdown of 25%, finish with a balance of $99.29, representing an ROI of -0.71%. Under threshold betting conditions SVM and LightGBM were the only two other models with negative ROIs of -0.28% and -1.06% respectively. All other models yielded positive results with the most successful models being XGBoost (12.10% ROI), KNN (11.18% ROI), and Voting (15.91% ROI). Apart from Naive Bayes, all other models, regardless of how successful, yielded much more consistent results (figure 8) in comparison to the flat betting results (figure 7).

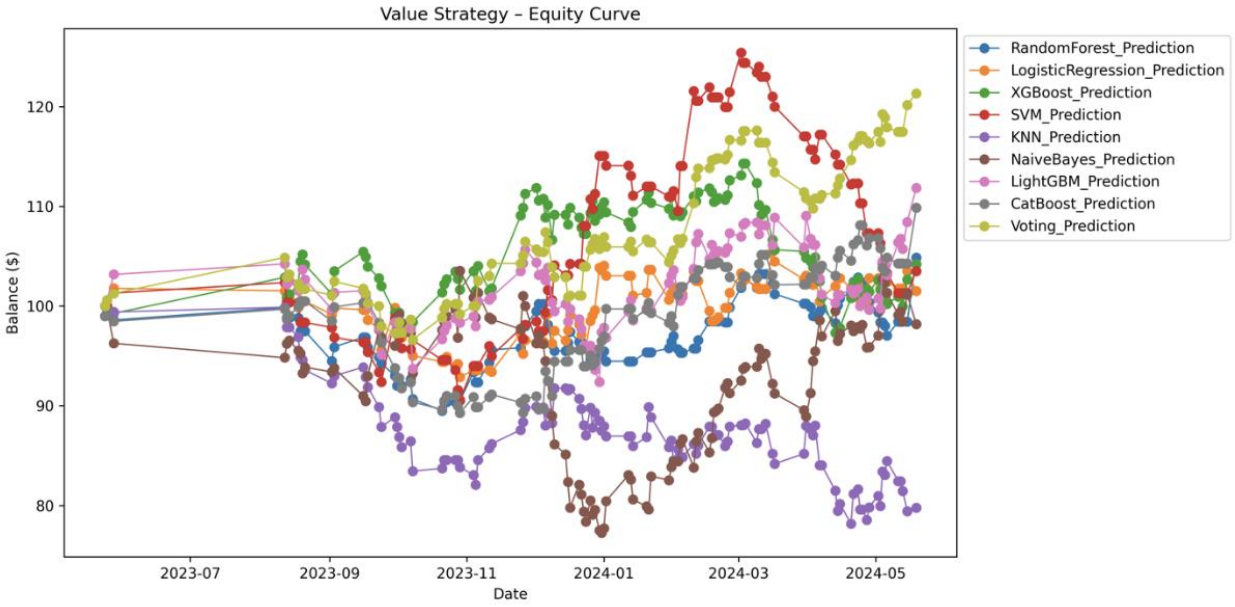


Figure 9: *Value Betting Equity Curve*

Value Betting involved placing bets only when the bookmaker odds exceeded the model-generated probabilities, targeting perceived mispricing by bookmakers. The same value wager placed for each mispricing means; this strategy is reliant on models consistently being able to provide more accurate odds than bookmakers. Other than the Voting model (21.32% ROI), LightGBM (11.84% ROI) and CatBoost (9.85% ROI) ended with the highest returns. As seen on the value betting equity curve (figure 9), the SVM model has the highest balance of >$120 over the period, however, finished with a closing balance of $103.50, displaying an inability to consistently provide more accurate odds than bookmakers over a longer period. Moreover, Naive Bayes also struggled under these conditions, however, managed to recover and salvage a closing balance of $98.19 (-1.81% ROI). This is also shown through Naive Bayes' max drawdown of

25% (Table 5). Lastly, the KNN model provided consistent negative results under the value betting conditions, finishing with a negative return of -20.22% and closing balance of $79.78, highlighting its difficulties in effectively identifying genuine betting value.

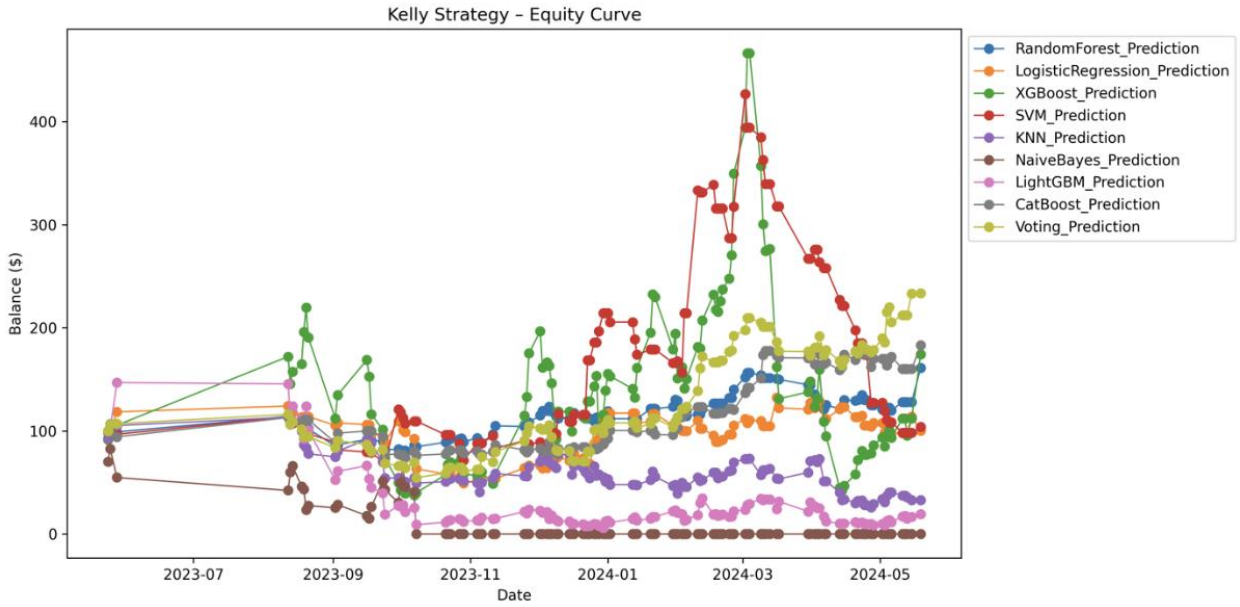


Figure 10: *Kelly Criterion Betting Equity Curve*

The Kelly Criterion strategy dynamically adjusted wager sizes according to each model's perceived betting advantage, meaning bets varied significantly depending on the calculated probabilities and bookmaker odds. The Kelly Criterion equity curve (figure 10) shows notable volatility, particularly evident in the XGBoost and SVM models, which experienced substantial gains, peaking sharply at over $400, before rapidly declining (91% and 77% max drawdown respectively). Naive Bayes and LightGBM experienced severe downturns, with Naive Bayes quickly using its entire bankroll at finishing with a balance of $0 after a less than half the period. While some models such as Random Forest, Logistic Regression and Voting provided more constant results, the nature of this betting strategy creates the possibility for drastic changes to amounts wagered and balance sizes throughout the simulation. The most successful models over the entire period under Kelly Criterion conditions proved to be CatBoost (83.15% ROI), XGBoost (74.20% ROI) and Voting (133.63% ROI), all with the most substantial returns of any models for any strategy.

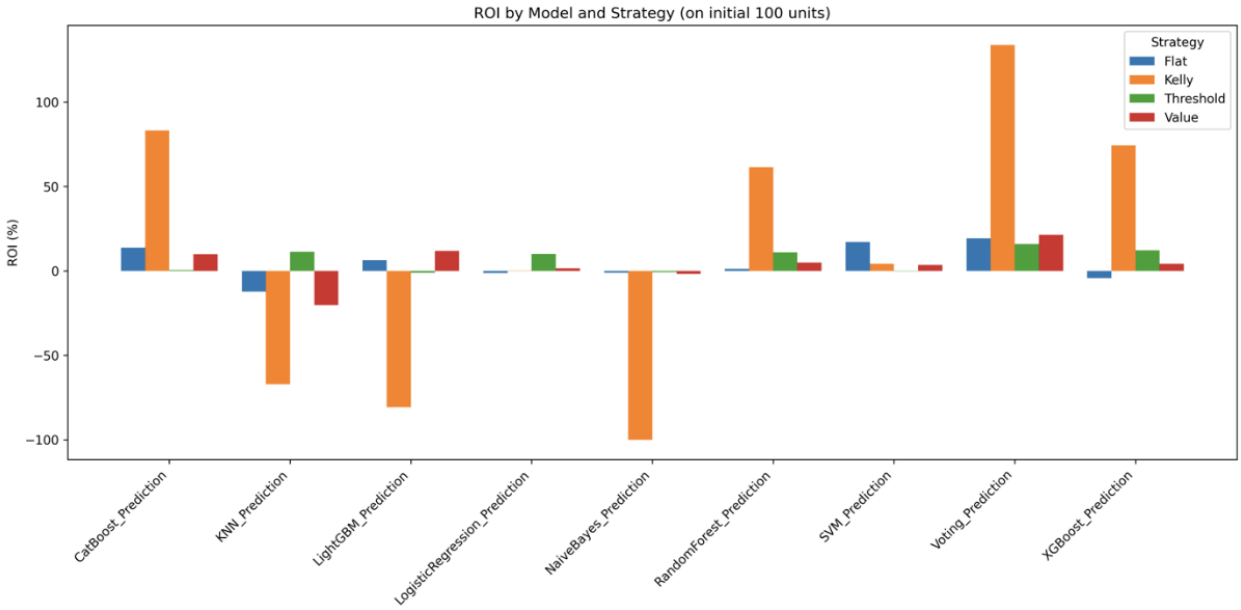


Figure 11: *ROI By Model Per Strategy*

Figure 11 demonstrates the ROI for each model at the end of the period, clearly showing the differences in model and strategy performance. In addition to the returns yielded, it is also important to evaluate overall betting activity, accuracy, and the number of wagers placed across strategies. The total amount staked varied significantly across strategies, reflecting their inherent differences. Flat Betting involved the simplest staking approach, with each model wagering a fixed total of $369 over the test period ($1 per game). Threshold Betting significantly reduced the number of wagers placed due to its selectivity, ranging from just 29 bets (SVM) to 364 bets (Naive Bayes). Value Betting involved a more moderate number of wagers, ranging from 131 bets (SVM) to 368 bets (Naive Bayes). The Kelly Criterion strategy showed the greatest variation in the total amount staked, from $669.71 with Naive Bayes up to $6,268.71 with XGBoost (Table 5), showing the high-risk nature of this method.

Flat Betting	Threshold Betting	Value Betting	Kelly Criterion
50.81%	70.89%	47.90%	46.67%

Table 6: *Average Win % Across all ML Models*

Looking at win percentages, Threshold Betting provided the highest accuracy rates overall, with Random Forest achieving an impressive 75%-win rate, the highest of any model across all strategies (Table 6). Logistic Regression (74.63%) and Voting (74.40%) also performed strongly under this strategy. The consistently high win percentages under Threshold Betting suggest that a confidence-based approach effectively identifies high-probability outcomes. Conversely, Value Betting and Kelly Criterion strategies showed lower overall win rates (Table 6) due to their reliance on odds discrepancies rather than outright prediction accuracy. However, an interesting note is that the Value Betting and Kelly Criterion strategies had <50% average win rates across models yet, still boast the models with higher ROIs compared to Flat and Threshold Betting, where the Win % are significantly higher across models.

Discussion

Performance Breakdown by Model

The results revealed substantial variation in the predictive performance of the machine learning models tested, highlighting important insights about the characteristics that made certain models more effective at predicting game outcomes. Among the tested algorithms, CatBoost and Random Forest notably excelled, achieving the highest accuracy (0.59 and 0.58, respectively) (Table 4). CatBoost's superior performance can primarily be attributed to its built-in handling of categorical variables using ordered target encoding, enabling robust probability estimates and limiting overfitting (Prokhorenkova et al., 2018). In the context of football betting, CatBoost's robust probability estimates likely allowed the model to consistently produce predictions closer to reality compared to bookmaker odds, providing tangible betting value. Random Forest, distinguished by the highest AUC (0.71), effectively modelled complex non-linear relationships without extensive tuning (Breiman, 2001), capturing nuanced interactions inherent in Premier League match data.

Logistic Regression and Naive Bayes performed comparably overall but showed contrasting strengths and weaknesses. Logistic Regression delivered clear, calibrated probability estimates beneficial for scenarios with stable patterns, yet struggled to predict less frequent outcomes, such

as draws, indicated by their moderate recall and F1-score (0.47 and 0.42 respectively). Naive Bayes, on the other hand, despite reasonable accuracy, had some notable issues with probabilistic calibration (Brier score of 0.94) due to its strong assumption of feature independence, a significant limitation given the correlated nature of sports data (James et al., 2021).

XGBoost and LightGBM demonstrated balanced predictive power, underpinned by their efficient, tree-based ensemble boosting approaches. Both models maintained competitive accuracy (0.56 and 0.54, respectively) and strong AUC scores (0.70). These algorithms effectively managed complex interactions and nonlinearities inherent in the dataset, enhanced by their optimization strategies specifically designed for structured data and their capacity to adaptively handle intricate relationships among predictors (Chen & Guestrin, 2016; Ke et al., 2017). Specifically, their ability to manage complex interactions allowed them to capture nuanced relationships between key football-specific factors, such as team form, and home advantage, enhancing their predictive effectiveness for Premier League outcomes.

Conversely, the KNN and SVM models displayed lower performances across most metrics (accuracy scores of 0.55 and 0.56, respectively). KNN, a non-parametric method heavily reliant on local similarities, struggled with the high-dimensional and noisy nature of football data, which reduced its effectiveness in generalizing across a broader set of matches (Cover & Hart, 1967). However, KNN typically can classify outcomes based on local similarity in historical patterns. Prior to observing the results of this exploration, we considered that this might be useful in football contexts, where matchups between similarly performing teams or recent form-based clusters can offer strong predictive cues. Unfortunately, upon observing the results yielded by this model's performance on the football data, we were quickly dissuaded of these notions.

SVM, while theoretically capable of handling intricate decision boundaries, suffered notably in precision (0.37), indicating challenges in class separation within the feature space. Its performance might suggest some difficulties in managing the inherent complexities and overlapping distributions typical in match prediction, potentially worsened by the complexities of calibrating its kernel parameters (Cortes & Vapnik, 1995).

The Voting classifier's powerful performance (accuracy 0.59, AUC 0.71) further emphasizes the effectiveness of ensemble methods. By combining diverse modelling perspectives, the Voting approach successfully mitigated individual model biases and capitalized on the unique strengths of each underlying classifier, thus enhancing overall robustness and reliability of predictions (Polikar, 2006). Because soft voting was used, it is comparable to assessing each model, which gives a probability (like 70% yes, 30% no), and you average these probabilities to choose the final prediction. Because the Voting classifier uses an ensemble method, it is unsurprising that this Model was one of the best overall.

While several models showed promising predictive capabilities, it is essential to maintain a cautious interpretation of these findings. Despite strong performances from algorithms like CatBoost, Random Forest, and the Voting classifier, none achieved overwhelming predictive dominance, indicating persistent limitations inherent in machine learning-based football outcome prediction. Models such as KNN and SVM underscore that general theoretical strengths do not always translate effectively into practice, particularly given the noisy, high-dimensional, and complex nature of sports data. Furthermore, the performance variations observed raise important considerations regarding model robustness, calibration accuracy, and reliance on historical data that may not fully account for dynamic real-world factors such as player form fluctuations or unexpected tactical changes. Thus, while ensemble methods offer a pragmatic solution by mitigating individual model shortcomings, their success should not overshadow the ongoing need for methodological improvements and cautious application in sports betting scenarios.

Feature importance

The dominance of corner-odds suggests that the betting market's expectation of set-piece frequency captures latent information about both teams' tactical profiles and match tempo. Meanwhile, recent form metrics, including goal-conversion efficiency and discipline ratios, provide a complementary, data-driven signal that is uncorrelated to raw odds. The variation in feature importance across model families highlights how different algorithms extract distinct facets of the data. Linear methods distil broad handicap and discipline trends, kernel-based and instance-based methods leverage fine-grained recent performance or specific over/under markets, and generative approaches rely on team identity priors. Remarkably, Naive Bayes even picks up

on the impact of red cards from prior matches. Taken together, these findings underscore the value of combining market-based signals with engineered form features in a heterogeneous ensemble to achieve robust match-outcome predictions (top feature importance rankings are provided in Appendix II).

Performance Breakdown by Strategy

The results from the betting simulations revealed interesting differences in the effectiveness of the four betting strategies used. Each strategy has a distinct philosophy toward managing uncertainty and utilizing predicted outcomes to generate a return.

Flat Betting provided a consistent, low-volatility baseline by applying a fixed stake to every prediction, regardless of confidence or implied value. This approach produced stable outcomes, with several models, such as Voting (19.19% ROI), SVM (17.11% ROI), and CatBoost (13.72% ROI), achieving very respectable profits. The strength of Flat Betting lies in its simplicity and even exposure across all predictions, which avoids overcommitting to any single game or overreacting to model confidence. Flat betting provides volume. The number of bets placed is higher than any other strategy, because a wager is placed on every game, regardless of certainty. Interestingly, if odds were completely *fair*, in terms of, $1 wagered equals $2 return (if won), for each bet, then models would only need to correctly predict more than 50% of the games to profit over the long term. However, this is not the case in reality, and where this strategy falls short is it lacks any mechanism to distinguish between strong and weak signals. It bets indiscriminately, which also includes lower-confidence predictions, thereby capping potential upside. Additionally, it fails to exploit the edge that may exist when model confidence or bookmaker mispricing is particularly strong, potentially leaving value on the table.

Threshold Betting introduced selectivity, only placing bets when the model's predicted probability exceeded a fixed confidence threshold (set at 60% in this exploration). This filtering mechanism reduced bet volume, ranging from 29 bets for SVM to 364 for Naive Bayes, and improved win rates across most models. The reason Naive bayes has an insignificant difference between the number of bets placed under the flat betting strategy compared to the threshold

strategy is its high confidence in all bets made. The strategy's success, shown by its 75% accuracy on average across models, can be attributed to concentrating wagers on higher-confidence predictions, improving hit-rates, and reducing exposure to less reliable outcomes. The top-performing models under this strategy, Voting (15.91% ROI), XGBoost (12.10%), and KNN (11.18%), highlight the power of this focused approach. However, Threshold Betting's effectiveness is sensitive to the chosen confidence level. A threshold that is too high results in too few bets, limiting ROI potential and increasing exposure to variance (a threshold that is too low reintroduces low-confidence noise). Finding the optimal threshold is therefore crucial and could be further refined in a future project through adaptive or model-specific calibration.

Value Betting performed best when models were able to consistently identify bookmaker mispricing, which are cases where the predicted win probability implied higher expected returns than the bookmaker odds offered. This strategy reflects a more nuanced, probability-driven approach and rewards well-calibrated models capable of recognizing inefficiencies in the market. Notably, the Voting model (21.32% ROI) and LightGBM (11.84% ROI) performed well under this strategy, illustrating that accurate probability estimation is essential to extract value. The lower average win rates observed under this strategy (47.9% across models) underline that success here depends not on raw accuracy, but on capturing mispriced high-odds outcomes. However, Value Betting assumes static bookmaker odds and does not account for line movement or bet timing, factors that could significantly impact real-world replicability. Furthermore, minor miscalibrations in model probabilities can quickly erode profitability, especially if the model consistently overestimates value in longer odds.

The Kelly Criterion strategy, which dynamically scales stake size based on the perceived edge, delivered the highest peak returns but also introduced significant volatility and risk. Models like Voting (133.63% ROI), CatBoost (83.15% ROI), and XGBoost (74.20% ROI) achieved the most impressive returns under this method. These outcomes reflect Kelly's theoretical strength when a model's edge is real and consistent, the strategy allocates capital efficiently to maximize upside. However, the same approach also amplifies the consequences of incorrect or overconfident predictions. This was evident in models like Naive Bayes and LightGBM, which lost their entire bankrolls (or close to it) under Kelly conditions. Naive Bayes, for example, was particularly

susceptible to over-betting due to poor probability calibration (as evidenced by its high Brier score of 0.94), leading to rapid drawdowns and eventual failure. These results illustrate that while the Kelly Criterion is a useful strategy for maximizing returns, it demands extremely accurate and stable probability estimates, a condition not all models are capable of meeting consistently. This model underscores that to achieve higher returns; there is an element of higher risk that must also be considered. In practical applications, a lower limit fractional Kelly or capped stake version might hold potential to mitigate these risks. This would involve limiting the proportion of the bankroll accessible at any given time, or below a certain threshold. This could help mitigate the downside risk, however, could also end up capping the potential up-side returns.

A key consideration across all strategies is the trade-off between risk and reward. Flat Betting favours capital preservation but lacks adaptability. Threshold Betting improves efficiency through selectivity but risks underutilization. Value Betting capitalizes on market inefficiencies but depends heavily on calibration. The Kelly Criterion promises maximum theoretical profit but only when probability estimates are both accurate and reliable. Importantly, real-world implementation would introduce further complexities including odds shifts, stake limits, and market liquidity, which were not modelled in these simulations.

When comparing the returns of machine learning-driven sports betting to traditional investment avenues, the contrast is striking. For instance, the S&P 500 has delivered an average annual return of approximately 10.13% since 1957 (Curvo, 2025), and the STOXX Europe 600, representing a broad spectrum of European companies, has demonstrated a compound annual growth rate (CAGR) of 7.94% from 1986 (Curvo, 2025). Contrasting these figures with our sports betting simulations, certain machine learning models, particularly when employing the Kelly Criterion strategy, achieved returns exceeding 130% ROI. While these returns are higher than traditional investments, they come with increased volatility and risk. The potential for significant gains is counterbalanced by the possibility of rapid losses, as evidenced by some models losing their entire bankrolls under certain strategies. Therefore, it is reasonable to infer, that while machine learning-driven sports betting can yield impressive short-term profits, the

associated risks and market limitations suggest that traditional investments may offer more stable and sustainable growth over the long term.

Real-World Limitations

While the results of this study demonstrate promising predictive accuracy and profitability under simulated conditions, applying these models and strategies in real-world betting markets presents several important limitations. These constraints must be carefully considered when assessing the practical viability of using machine learning to bet on future Premier League seasons.

Odds Availability and Line Movement

One of the most immediate limitations is the assumption of static odds. In this exploration, bookmaker odds prior to kick-off were taken and used as fixed inputs for betting simulations. In reality, odds can fluctuate rapidly due to market activity, team news, or external factors. If a model identifies a value opportunity at odds of 3.00, those odds may drop to 2.40 before a user has time to place a wager, effectively erasing the edge. This phenomenon, known as line movement, introduces execution risk. A successful real-world implementation would require real-time odds monitoring and fast bet placement, ideally through automation or APIs, to minimize slippage. Without such infrastructure, the model's theoretical profitability may be significantly reduced.

Injury, Lineup, and External Information Gaps

The predictive models used in this project rely solely on pre-match statistics and rolling averages with varying window sizes (1, 3, 5, and 10). They do not incorporate real-time information such as player injuries, starting lineups, tactical shifts, or motivational factors, variables that bookmakers actively price into their odds. For instance, a key striker being ruled out an hour before kick-off can meaningfully shift the actual probability of an outcome, but unless that data is incorporated, the model continues to operate on out-of-date assumptions. This asymmetry puts model-driven bettors at a disadvantage, as bookmakers and professional groups often have faster access to news and thereby adjust their odds accordingly. Addressing this gap would require

integrating some structured and unstructured data feeds, such as injury reports, lineup announcements, or social media sentiment, which may increase model complexity and significantly increase data costs. Again, access to information regardless may still be slower than bookmakers, meaning this added complexity might not yield the desired advantage.

Market Reaction to Volume and Liquidity Constraints

Another practical limitation relates to how bookmakers respond to large or consistent betting volume. The simulations assume that any bet, regardless of size, can be placed at the posted odds. In practice, bookmakers, such as Bet365, can react to sharp betting activity. Bookmakers may lower odds in response to a single large wager or a series of bets from the same source, especially if the bets consistently target mispriced outcomes. This is a major concern for strategies like value betting or Kelly Criterion, which focus bets where the model detects the greatest edge. If a model consistently identifies high-value situations, repeated wagering could alter market prices or lead to stake limitations. Alternatively, bookmakers can cash-out betting accounts and ban users from the platform. In practice, this means that the more effective a model becomes, the harder it is to scale without impacting the market it is trying to exploit.

Betting Account Restrictions and Sustainability

The most underappreciated limitation is the viability of deploying profitable strategies through traditional betting accounts. Bookmakers actively monitor account performance and frequently impose restrictions or outright bans on users who consistently win, place value bets, or display signs of algorithmic behaviour. This is especially problematic for strategies such as value betting and Kelly criterion betting, which could signal sharp activity in the eyes of the bookmakers. Even modest success over time can lead to stake limits or market access denial, effectively capping potential profitability. To remain viable, bettors may need to spread activity across multiple accounts or shift to betting exchanges, both of which introduce additional logistical and legal challenges. While not explicitly illegal in the Danish betting market, algorithmic betting would violate most bookmakers' (such as Bet365) terms and conditions which are grounds for a ban. Additionally, spreading betting activity across several accounts can prove difficult as Danish betting profiles are typically linked to MitID, the Danish digital ID, meaning a single

user cannot own multiple accounts. Using these models and strategies would therefore be less legally complex if implemented in countries with less gambling regulation.

Future Research & Exploration

While this project demonstrates the potential for machine learning models to generate profit in Premier League betting markets, several promising avenues remain for further exploration. These directions could improve predictive performance, enhance betting strategies, and bring the methodology closer to real-world deployment.

In future work, strategy selection could benefit from personalization or hybrid approaches. For example, combining Threshold and Value strategies, only betting on high-confidence predictions that also offer value, might yield improved returns. Adaptive thresholding based on model-specific calibration curves could further optimize betting frequency. Finally, applying volatility-adjusted staking could allow for more robust deployment of dynamic staking without exposing the bankroll to extreme swings. Overall, while no single strategy guarantees success, the evidence suggests that aligning betting behaviour with model strengths, particularly in probability estimation, is essential to maximizing returns.

It is important to note that the machine learning models used in this exploration were not tuned specifically for each individual betting strategy. Instead, they were optimized primarily for flat betting to maximise ROI. The different betting strategies were then overlaid on top of these fixed model predictions during simulation. While this allowed for a clean comparison of strategies using a common predictive baseline, it also means that each model's performance under a given strategy may not reflect its full potential.

A promising direction for future research would be to develop machine learning models specifically for betting strategy optimization itself. Rather than using ML solely for outcome prediction, one could train a separate model, such as a regression algorithm, to directly learn the optimal betting parameters (e.g., thresholds, stake sizing rules) that maximize ROI. This meta-model would take as input the predicted probabilities and contextual features and output the most

profitable action. By shifting the learning objective from classification to profit optimization, this approach could dynamically adapt to market conditions and model behaviour, representing a powerful evolution of data-driven sports betting.

One clear area for future research is the use of multi-bookmaker environments. This study used a single set of odds (primarily from Bet365) to place bets; however, it did use multiple odds from different bookmakers as features. By comparing lines across European and Asian bookmaker markets, researchers could detect discrepancies and identify arbitrage or value opportunities. Asian bookmakers, particularly Pinnacle and SBOBET, are known for having more efficient markets and narrower margins and often serve as market makers for the global betting ecosystem. There is also increasing evidence that major Asian bookmakers deploy machine learning models to adjust odds dynamically based on betting volume and historical trends (Playlogiq, 2023). A compelling future approach could involve using a lightweight model to systematically compare odds from Asian and European bookmakers in real time, flagging bets where one market significantly deviates from the other. This could serve as a powerful signal for value betting or odds inefficiency, especially if one market is already using machine learning to create more accurate odds. By leveraging line discrepancies and market inefficiencies, this method would move beyond outcome prediction and into comparative pricing strategy.

A final opportunity lies in exploring more granular market types. While this study focused on full-time match results (win/draw/loss), other markets may offer better profit potential or be less efficient. Examples include over/under goals, Asian handicap spreads, halftime scores, total corners, and player-specific props like goal scorers or cards. These markets often receive less attention from sharp bettors and are sometimes priced more loosely by bookmakers. From a modelling perspective, many of these outcomes can be reframed as either classification or regression tasks depending on the available data. For instance, predicting expected goals and then mapping those to over/under thresholds, or forecasting expected corner counts and betting on associated ranges. Future work could collect more granular game statistics or player-level data to power such models, potentially opening entirely new value channels.

Conclusion

This research aimed to evaluate how effectively machine learning models can generate sustainable profits within the English Premier League betting market. Specifically, we explored the intersection of predictive analytics and betting strategies, seeking not only predictive accuracy but also tangible financial outcomes. The guiding research question was:

"*How effective are machine learning models in generating sustainable profits within the football betting market?*"

Our findings demonstrate that, indeed, machine learning models, when strategically combined with targeted betting strategies, have significant potential to yield consistent profitability. The ensemble Voting model emerged as the standout performer, notably achieving impressive returns across multiple strategies, such as the Kelly Criterion (133.63% ROI), Flat Betting (19.19% ROI), Threshold Betting (15.91% ROI), and Value Betting (21.32% ROI). CatBoost and Random Forest models also consistently provided strong returns, underscoring their ability to effectively capture the complex, non-linear dynamics inherent in Premier League football data.

While the performance metrics used were robust indicators of a model's classification capabilities, our study importantly highlighted that profitability in sports betting markets is not only dependent on these predictive metrics. The Kelly Criterion strategy, for instance, leveraged probability calibration and betting edge to maximize returns, highlighting the most significant potential for profit but also introducing considerable risk and volatility. Conversely, Threshold Betting, focusing on high-confidence predictions, demonstrated that enhanced selectivity can stabilize returns and improve win rates, albeit with reduced betting volume.

Interestingly, despite the high accuracy of certain models, such as CatBoost (59%), it became evident that consistent profit generation depended on how betting strategies leveraged these predictions. Models such as Naive Bayes and KNN struggled significantly in terms of probability calibration and betting returns, highlighting the critical importance of precise probability estimation.

Furthermore, this exploration identified some important practical limitations that must be addressed before deploying such models and strategies in real-world scenarios and upcoming seasons. Issues such as the dynamic nature of bookmaker odds, execution timing, data accessibility (player injuries, team news), and how bookmaker-imposed restrictions significantly complicate real-world implementation. Addressing these challenges in future research, potentially through automated bet placement systems, real-time data integration, and cross-market odds comparisons, could further enhance the practical utility and profitability of machine learning in sports betting.

Future research could beneficially explore the potential of hybrid or adaptive betting strategies. Combining Threshold and Value Betting strategies, or utilizing stake capped Kelly Criterion adjustments, might offer more balanced approaches, optimizing risk vs. reward trade-offs. Additionally, expanding analyses into other betting types, such as goal totals, Asian handicaps, or player-specific propositions, could uncover further profitable avenues less efficiently priced by bookmakers. As models and strategies become more refined, contrasting and comparing to other investment types becomes logical, if the objective is maximizing profitability.

This study therefore provides evidence that machine learning, combined with strategic betting approaches, can outperform traditional betting methods in the English Premier League market during our test period of the 23/24 season. In direct response to our research question, we conclude that machine learning models can be highly effective in generating sustainable profits within the football betting market albeit with higher risk level than traditional investments. Our findings confirm that, while challenges remain, the intersection of predictive analytics and strategic betting holds considerable promise for sustainable profitability in sports betting markets.

Bibliography

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). Artificial Intelligence Review, 54(3), 1937–1967. https://repositorio.uam.es/bitstream/handle/10486/714798/7912259_ps.pdf?sequence=1&isAllowed=y

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A., Deng, D., & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, *13*(2). https://doi.org/10.1002/widm.1484

Bonn, K. (2025, April 20). Worst Manchester United seasons in Premier League history: Ruben Amorim on course to break Erik ten Hag records. *Sporting News United Kingdom*. https://www.sportingnews.com/uk/football/news/manchester-united-premier-league-worst-seasons-history-rank/jwcd9h6rsqffev6br7yu4bzw

Breiman, L. (2001). Machine Learning, 45(1), 5–32. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Bunker, R., Yeung, C., & Fujii, K. (2024, March 12). *Machine learning for soccer match result prediction*. arXiv.org. https://arxiv.org/abs/2403.07669

Chen, T., & Guestrin, C. (2016). Proceedings of ACM SIGKDD 2016, 785–794. https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf

Constantinou, A., & Fenton, N. (2013). Profiting from arbitrage and odds biases of the European football gambling market. Journal of Gambling Business and Economics, 7(2), 41–70. http://constantinou.info/downloads/papers/evidenceofinefficiency.pdf

Cortes, C., & Vapnik, V. (1995). Machine Learning, 20(3), 273–297. https://web.engr.oregonstate.edu/~huanlian/teaching/ML/2019spring/extra/svn-1995.pdf

Cover, T. M., & Hart, P. E. (1967). IEEE Transactions on Information Theory, 13(1), 21–27. https://isl.stanford.edu/~cover/papers/transIT/0021cove.pdf

Curvo. (2025). *Historical performance*. Curvo. https://curvo.eu/backtest/en/compare

EGBA. (2025). European Gambling Market. https://www.egba.eu/eu-market/

Elo, A. E. (1978). *The rating of chessplayers: Past and present*. Arco Publishing.

England football results betting odds | Premiership results & betting Odds. (n.d.). https://www.football-data.co.uk/englandm.php

Fama, E. F. (1970). Efficient Capital Markets: A review of theory and Empirical work. *The Journal of Finance*, *25*(2), 383. https://doi.org/10.2307/2325486

Federazione Italiana Giuoco Calcio, & AREL - Agenzia di Ricerche e Legislazione, & PwC. (August 6, 2024). Share of sports betting handle coming from bets on soccer worldwide in 2023, by continent [Graph]. In Statista. Retrieved May 06, 2025, from https://www.statista.com/statistics/1534841/sports-betting-handle-soccer-worldwide/

Forrest, D., & Simmons, R. (2001). Globalisation and efficiency in the fixed-odds soccer betting market (Tech. Rep.). University of Salford. https://www.yumpu.com/en/document/view/17554944/globalisation-and-efficiency-in-the-fixed-odds-soccer-betting-istituti

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), 1–37. https://doi.org/10.1145/2523813

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer. https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf

James, G., et al. (2021). An Introduction to Statistical Learning. Springer. https://www.casact.org/sites/default/files/2022-12/James-G.-et-al.-2nd-edition-Springer-2021.pdf

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Kelly Jr, J. L. (2011). A new interpretation of information rate. In The Kelly Capital Growth Investment Criterion: Theory and Practice (pp. 25–34). World Scientific. (Original work published 1956) https://www.worldscientific.com/doi/abs/10.1142/9789814293501_0003?srsltid=AfmBOorT6jymOAs9EJpfyQfsxlCdpHznKptm-DaxkKb3YtMx7wV2EZPR

Leicester's Premier League heroes. (n.d.). ESPN.com. https://www.espn.com/espn/feature/story/_/id/15386257/the-leicester-city-heroes-won-premier-league

Manchester United History. (n.d.). https://ir.manutd.com/company-information/history.aspx#:~:text=Since%201992%2C%20we%20have%20won,most%20successful%20clubs%20in%20England.

Newall, P. W. (2015). How bookies make your money. Judgment and Decision Making, 10(3), 225–231. https://www.researchgate.net/publication/279997433_How_bookies_make_your_money

PlayLogiq. (2023, November 22). How machine learning could change the sports betting industry. Logiq. https://playlogiq.com/igaming/how-machine-learning-could-change-the-sports-betting-industry/

Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3), 21–45. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1688199

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6639–6649. https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. Journal of the Royal Statistical Society: Series D (The Statistician), 49(3), 399–418. https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9884.00243

Stübinger, J., Mangold, B., & Knoll, J. (2019). Machine Learning in football Betting: Prediction of match results based on player characteristics. *Applied Sciences*, *10*(1), 46. https://doi.org/10.3390/app10010046

The football landscape – The Vision 2020-2023 | FIFA Publications. (n.d.). FIFA Publications. https://publications.fifa.com/en/vision-report-2021/the-football-landscape/

Thorp, E. O. (2006). The Kelly criterion in blackjack, sports betting, and the stock market. In Handbook of Asset and Liability Management (pp. 385–428). Elsevier. https://wayback.archive-it.org/5456/20240920160238/https://www.eecs.harvard.edu/cs286r/courses/fall12/papers/Thorpe_KellyCriterion2007.pdf

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS ONE*, *9*(1), e84217. https://doi.org/10.1371/journal.pone.0084217

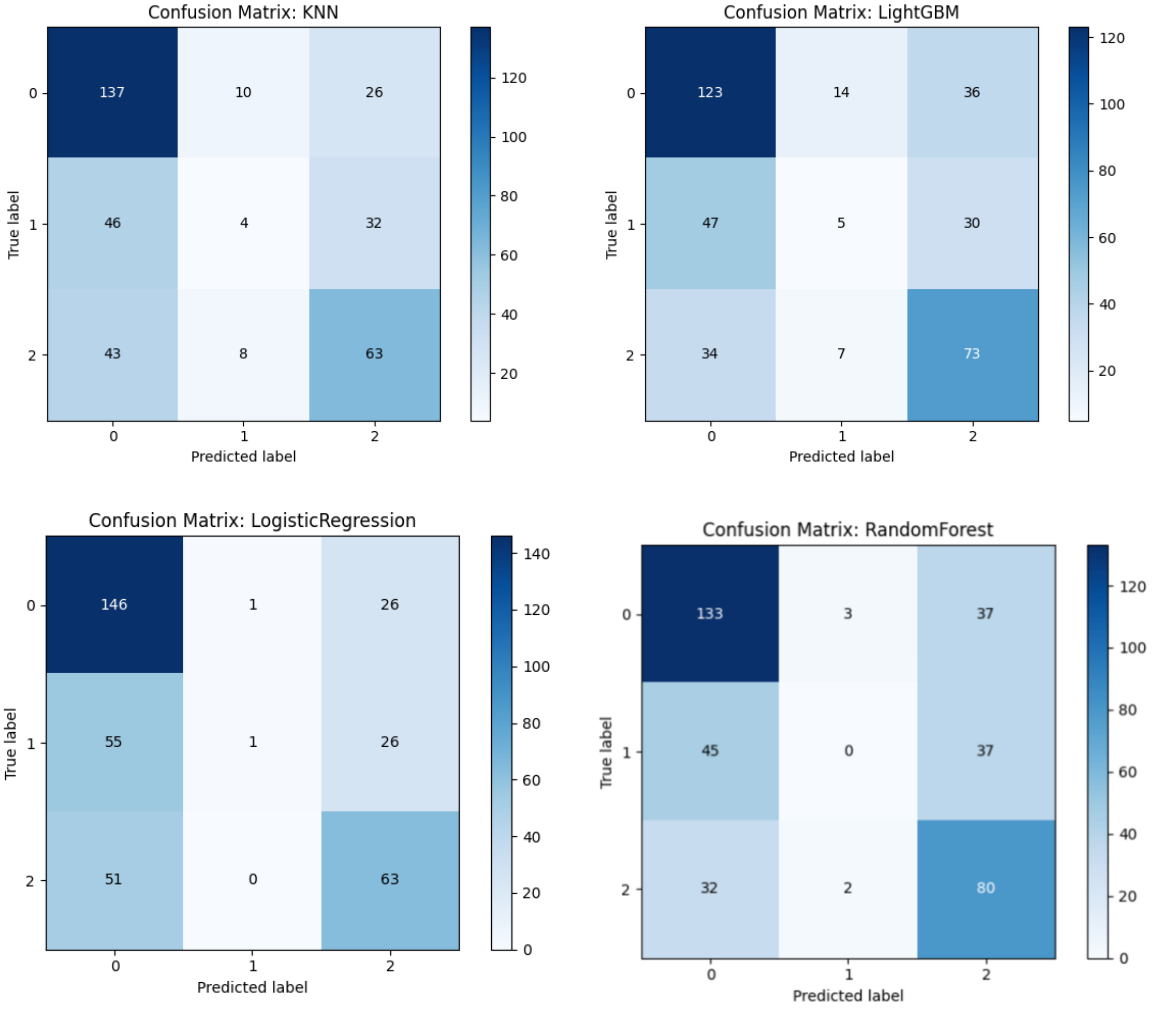
Vlastakis, N., Dotsis, G., & Markellos, R. N. (2006). Beating the odds: Arbitrage and winning strategies in the football betting market. Applied Financial Economics, 19(3), 183–197. https://ideas.repec.org/a/jof/jforec/v28y2009i5p426-444.html

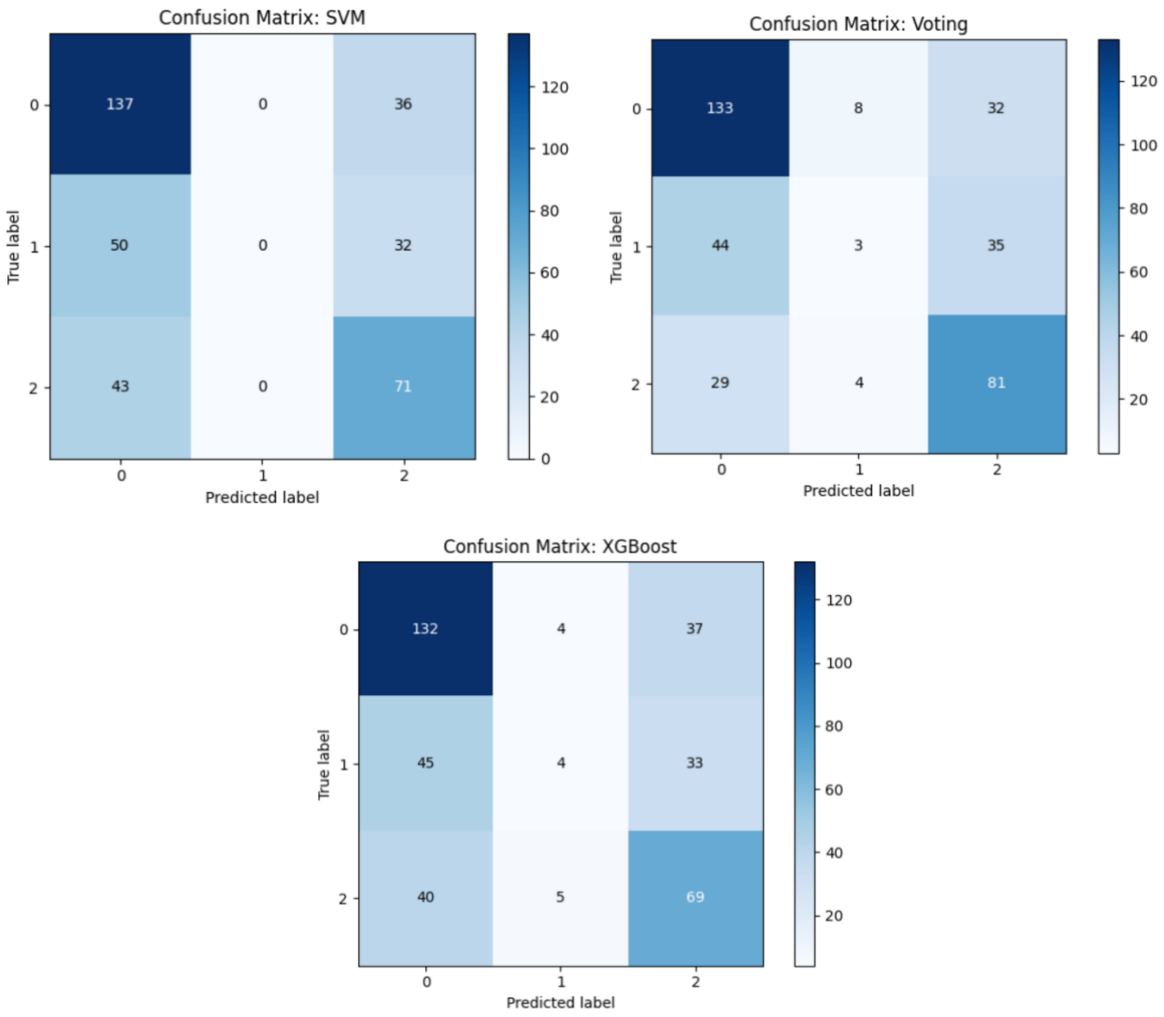
Wijk, D. van. (2021). Beating the Bookmakers using Machine Learning (Master's thesis, Erasmus University Rotterdam). https://thesis.eur.nl/pub/59277

Yeung, C., Bunker, R., Umemoto, R., & Fujii, K. (2024). Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees. *Machine Learning*, *113*(10), 7541–7564. https://doi.org/10.1007/s10994-024-06608-w

Appendix

Appendix I: Confusion matrices for remaining models




Confusion Matrix: SVM
True label
Predicted label
137
0
36
50
0
32
43
0
71
Confusion Matrix: Voting
True label
Predicted label
133
8
32
44
3
35
29
4
81
Confusion Matrix: XGBoost
True label
Predicted label
132
4
37
45
4
33
40
5
69

Appendix II: Feature importance of the top 5 features for each model

RandomForest	**LogisticRegression**	**XGBoost**
PSCA: **0.0129**	away_AF_last1: **0.0722**	PSCH: **0.0556**
MaxCH: **0.0118**	AHCh: **0.0676**	VCCH: **0.0447**
AvgA: **0.0106**	away_AR_last1: **0.0607**	MaxCH: **0.0260**
AvgCA: **0.0106**	PCAHA: **0.0543**	AvgCA: **0.0142**
B365CA: **0.0103**	away_foul_to_card_ratio_last3: **0.0511**	WHCH: **0.0135**

SVM	**KNN**	**NaiveBayes**
home_HY_last10: **0.0089**	home_FTHG_last3: **0.0057**	away_AR_last1: **0.0100**
away_foul_to_card_ratio_last3: **0.0076**	Max_less2.5: **0.0043**	Sheffield United: **0.0057**
home_HS_last5: **0.0049**	VCCD: **0.0027**	Brighton (Home): **0.0054**
home_HS_last10: **0.0041**	Avg_less2.5: **0.0027**	away_AR_last3: **0.0051**
home_HR_last5: **0.0024**	home_HST_last3: **0.0024**	Brighton (Away): **0.0049**

LightGBM	**CatBoost**	**Voting**
PSCH: **0.0389**	home_goal_conversion_last1: **0.0188**	VCCH: **0.0146**
MaxCH: **0.0344**	home_foul_to_card_ratio_last1: **0.0173**	PC_less2.5: **0.0127**
VCCH: **0.0268**	home_foul_to_card_ratio_last10: **0.0172**	away_foul_to_card_ratio: **0.0108**
away_goal_conversion: **0.0230**	home_foul_to_card_ratio_last5: **0.0170**	PSCA: **0.0106**
away_foul_to_card_ratio: **0.0221**	away_foul_to_card_ratio_last10: **0.0155**	home_goal_conversion_last5: **0.0098**