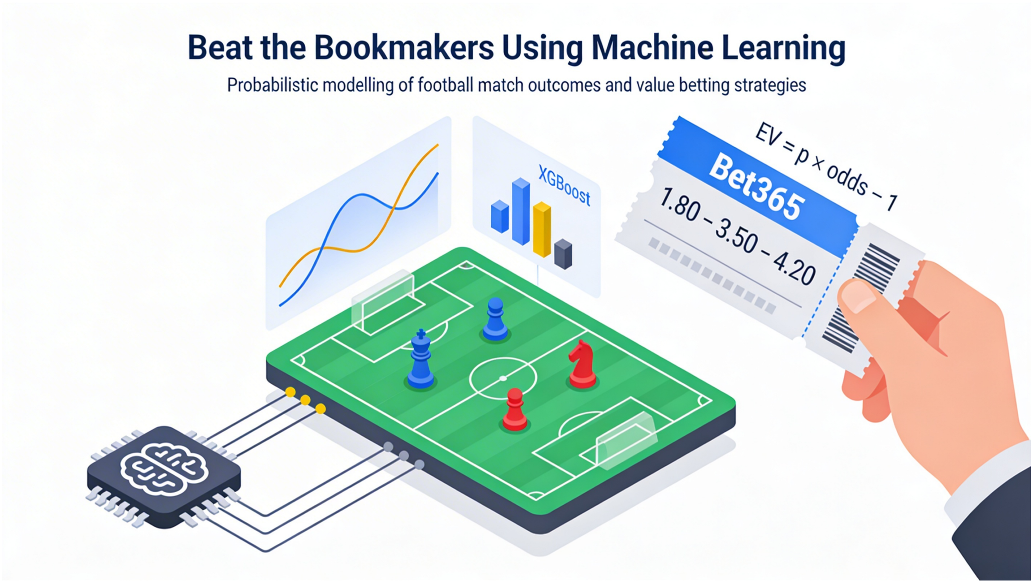
Beat the Bookmakers Using Machine Learning

Charles Defosseux   Alessandro Arensberg

February 15, 2026



Abstract

This project investigates whether a supervised machine learning model can identify profitable football bets by exploiting small inefficiencies in bookmaker odds. Using the European Soccer Database (Premier League and Ligue 1, 2008-2016) and Bet365 odds, we build match-level features, train probabilistic classifiers (multinomial logistic regression, gradient boosting), calibrate predicted probabilities, and transform them into expected values and betting strategies. Despite achieving reasonable predictive performance (test log-loss around 1.0 and accuracy near 50%), all tested strategies – including conservative value betting rules and a convex portfolio optimisation – yield strongly negative simulated returns. The study highlights the gap between statistical performance and economic profitability and illustrates the difficulty of beating an efficient betting market with historical, publicly available data.

Contents

1 Introduction

Sports betting is a large and highly competitive market in which bookmakers set odds that implicitly encode probabilities of different outcomes. In theory, a bettor who can estimate match outcome probabilities more accurately than the market can identify *value bets*, i.e. wagers with positive expected value (EV), and generate long-run profits.

The goal of this project is to assess whether a machine learning (ML) model trained on historical football data and bookmaker odds can systematically find such value bets. We focus on two major European leagues (English Premier League and French Ligue 1) over eight seasons, using the European Soccer Database from Kaggle and Bet365 odds. The core questions are:

- Can we build a model that predicts match outcomes $H/D/A$ with competitive probabilistic accuracy?
- When these probabilities are turned into betting decisions, does the strategy exhibit a positive return on investment (ROI) on historical data?

From a business perspective, the project evaluates the feasibility of a data-driven betting strategy under realistic constraints, and demonstrates how to connect ML metrics (log-loss, accuracy) to financial metrics (EV, ROI, drawdown).

2 Data and Business Problem

2.1 Data sources and schema

We use the *European Soccer Database* (Kaggle) which provides relational tables for matches, leagues, teams, players, team attributes, and bookmaker odds.

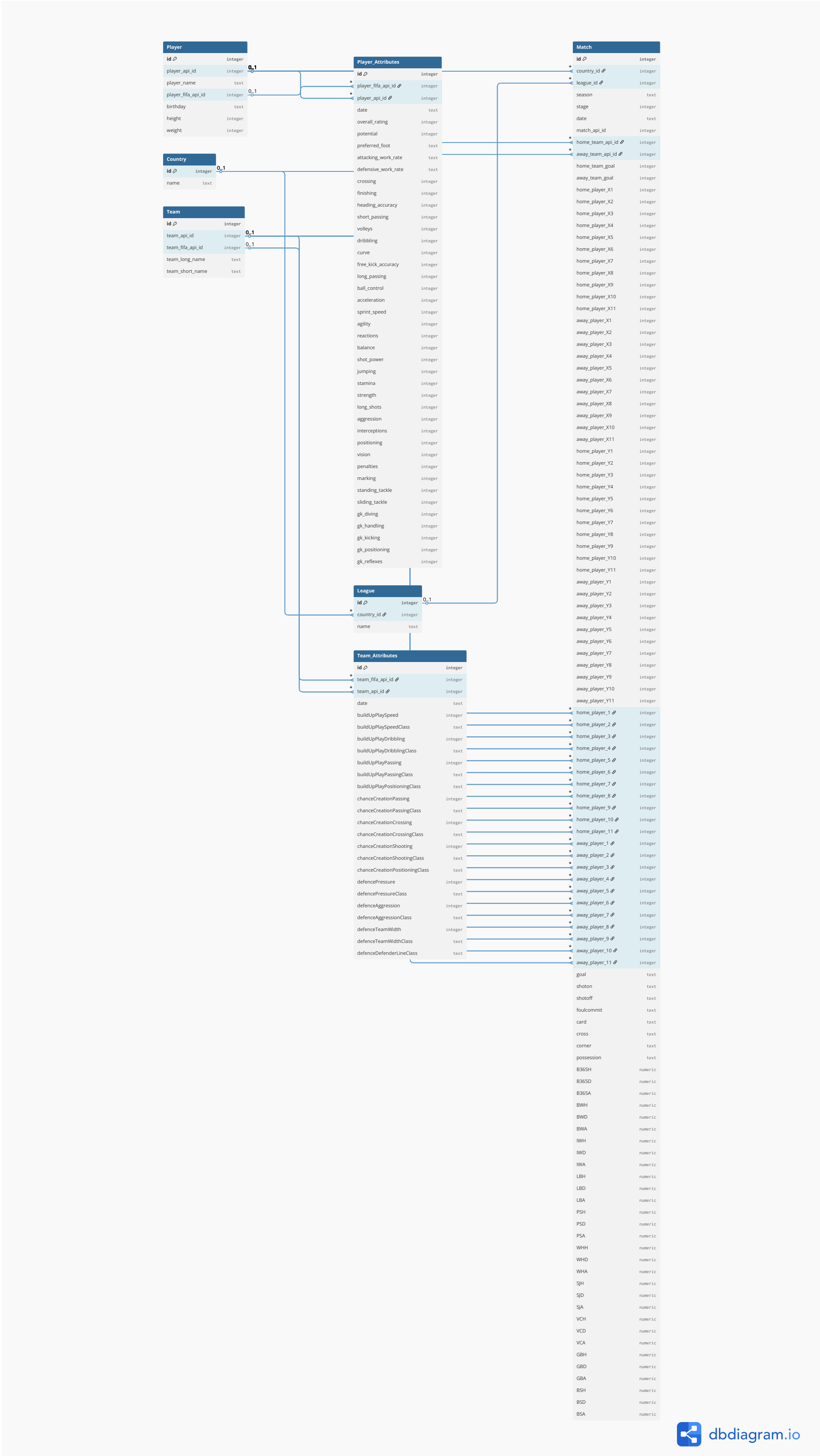


Figure 1: Simplified database schema (European Soccer Database).

We restrict the scope to:

- **Leagues**: English Premier League and French Ligue 1.
- **Seasons**: 2008/2009 to 2015/2016.
- **Bookmaker**: Bet365 home/draw/away odds (`B365H`, `B365D`, `B365A`).

After filtering and removing matches with missing Bet365 odds, we obtain a clean match-level dataset summarised in Table 1.

Table 1: Dataset overview: leagues, seasons and outcome frequencies.

league$_{name}$	n$_{matches}$	season$_{min}$	season$_{max}$	home$_{win_pct}$	draw$_{pct}$	away$_{win_pct}$
England Premier League	3040	2008/2009	2015/2016	45.7	25.8	28.5
France Ligue 1	3036	2008/2009	2015/2016	44.7	28.3	27.1

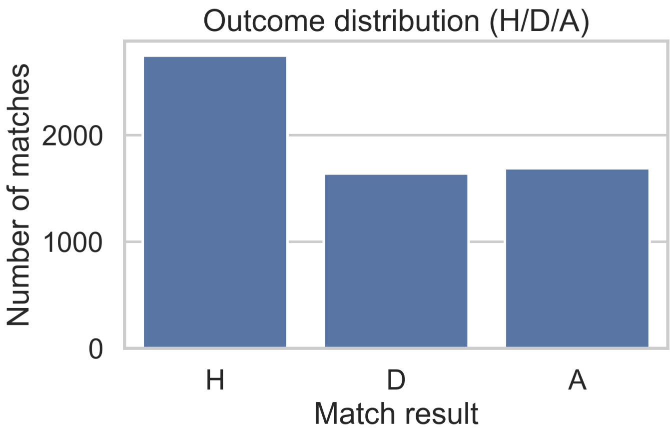


Figure 2: Distribution of match results (home win, draw, away win).

2.2 Business problem formulation

From a business standpoint, the problem is to design a betting strategy that:

1. Estimates for each upcoming match the probabilities $P(H)$, $P(D)$, $P(A)$.
2. Compares these probabilities to Bet365 implicit probabilities.
3. Places bets only when the expected value is positive and sufficiently high to justify risk.

Formally, for a given outcome $o \in \{H, D, A\}$ and match t, with model probability $p_{o,t}$ and bookmaker odds $c_{o,t}$, the expected value of a unit stake is:

$$\mathrm{EV}_{o,t} = p_{o,t} \cdot c_{o,t} - 1.$$

A naive value betting strategy would bet whenever $\mathrm{EV}_{o,t} > 0$. The central question is whether any such strategy, based on model probabilities, can deliver a positive ROI on historical data.

3 Feature Engineering

3.1 Market probabilities from odds

For each match, we use the Bet365 odds `B365H`, `B365D`, `B365A`. Naive implied probabilities are 1/odds, but include bookmaker margin. We compute margin-adjusted probabilities:

$$q_H = \frac{1/\texttt{B365H}}{1/\texttt{B365H} + 1/\texttt{B365D} + 1/\texttt{B365A}}, \quad \text{and similarly for } q_D, q_A.$$

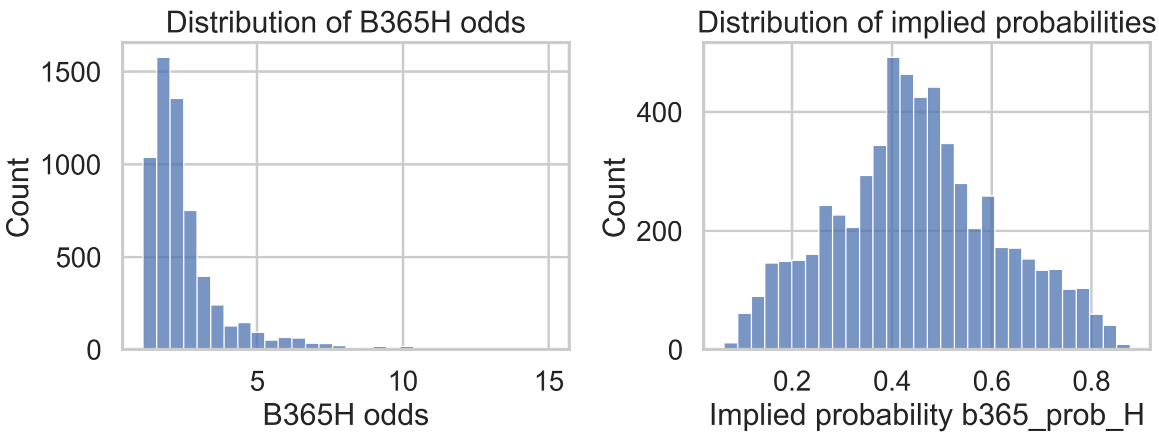


Figure 3: Distribution of Bet365 home odds (`B365H`) and implied probabilities (`b365_prob_H`).

3.2 Team form features

We construct rolling form features using the last $N = 5$ matches for each team, without temporal leakage (only past matches are considered). For each team and match, we compute:

- average goals scored and conceded,
- average points (3/1/0),
- win rate.

These are computed separately for the home and away team, as detailed in Table 2.

Table 2: Form features computed on the last $N = 5$ matches.

feature	description
home$_roll_goals_for$	Avg goals scored (home) over last N matches
home$_roll_goals_against$	Avg goals conceded (home) over last N matches
home$_roll_win_rate$	Win rate (home) over last N matches
away$_roll_goals_for$	Avg goals scored (away) over last N matches
away$_roll_goals_against$	Avg goals conceded (away) over last N matches
away$_roll_win_rate$	Win rate (away) over last N matches

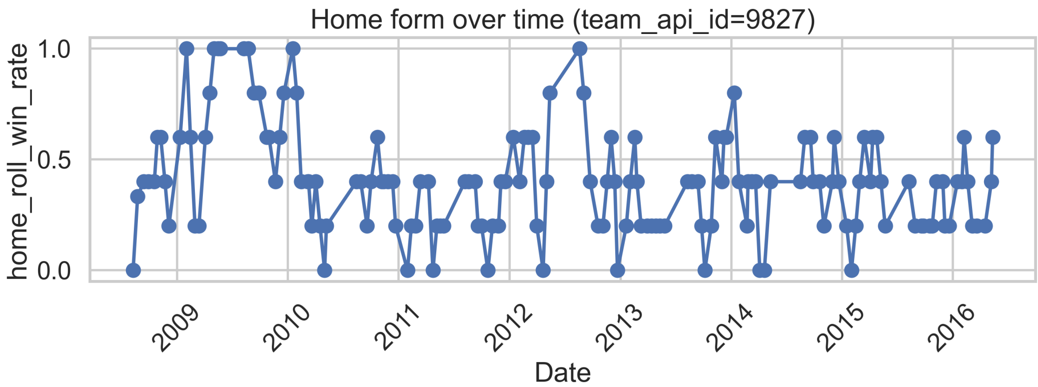


Figure 4: Example of home rolling win rate over time for a given team.

3.3 Team attributes

We augment each match with FIFA-style team attributes from the `Team_Attributes` table. For each team and match date, we select the most recent attribute record with date less than or equal to the match date. We then create:

- home and away attributes (e.g. `home_buildUpPlaySpeed`),
- differences home-minus-away (e.g. `diff_buildUpPlaySpeed`).

Table 3: Team attributes used as features.

attribute	description
buildUpPlaySpeed	Build-up play speed
buildUpPlayPassing	Build-up play passing quality
chanceCreationPassing	Chance creation via passing
chanceCreationCrossing	Chance creation via crossing
chanceCreationShooting	Chance creation via shooting
defencePressure	Defensive pressure intensity
defenceAggression	Defensive aggression
defenceTeamWidth	Defensive team width

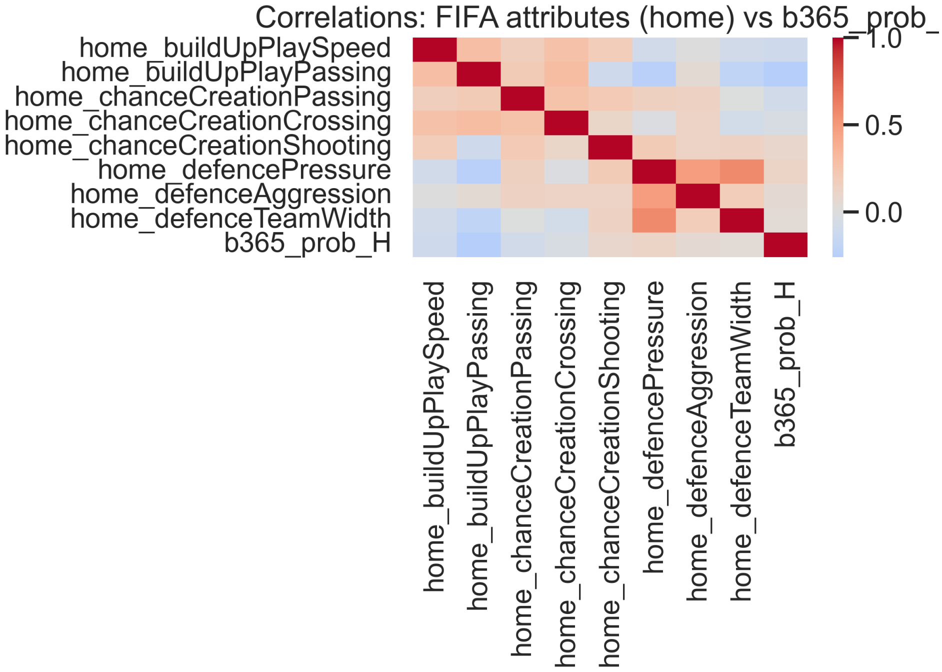


Figure 5: Correlation between home team attributes and market probability of home win.

4 Modelling Approach

4.1 Train/validation/test split

We adopt a chronological split based on season start year to mimic a realistic deployment scenario:

- **Train**: seasons up to 2011/2012,
- **Validation**: 2012/2013,
- **Test**: 2013/2014 and later.

Baselines include:

- always predicting a home win,
- predicting the most probable outcome according to Bet365 implied probabilities.

Table 4: Baseline models on the test set.

model	$\log_{l}oss_{t}est$	$\text{accuracy}_{t}est$
Always Home (naive)	NA	0.4462
Market (Bet365 probs)	0.9851	0.5287

4.2 Multinomial logistic regression

As a first model, we train a multinomial logistic regression with L2 regularisation. Input features include:

- Bet365 odds and implied probabilities,
- home and away form features,
- home and away team attributes and their differences.

Continuous features are standardised, and missing values are imputed by the median.

On the validation and test sets, the model achieves:

- validation log-loss ≈ 0.9935, test log-loss ≈ 0.9984,
- validation accuracy ≈ 0.51, test accuracy ≈ 0.52.

Table 5: Performance of multinomial logistic regression.

model	$\log_l oss_v al$	$\log_l oss_t est$	accuracy$_v al$	accuracy$_t est$
Logistic Regression (multinomial)	0.9935	0.9984	0.5105	0.5195

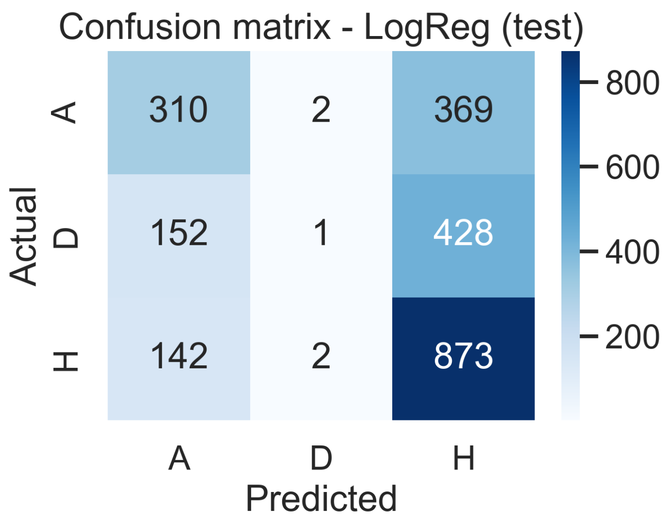


Figure 6: Confusion matrix for logistic regression on the test set.

4.3 Gradient boosting models (XGBoost)

To increase expressive power, we train gradient boosted trees (XGBoost) with multi-class soft-prob objective. We compare two feature sets:

1. **Without odds**: only form and team attribute features.
2. **With odds**: adding Bet365 odds and implied probabilities.

On the test set:

- XGBoost without odds: log-loss ≈ 1.0750, accuracy ≈ 0.4546.
- XGBoost with odds: log-loss ≈ 1.0319, accuracy ≈ 0.4976.

Table 6: Comparison of model performance on the test set.

model	$\log_l oss_t est$	accuracy$_t est$
LogReg + odds + form + attributes	0.9984	0.5195
XGBoost no odds	1.075	0.4546
XGBoost with odds	1.0319	0.4976

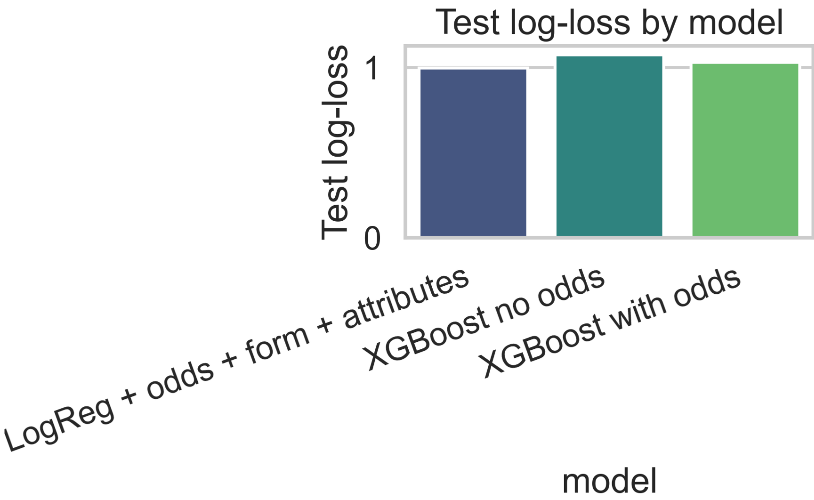


Figure 7: Test log-loss by model (lower is better).

The large gap between XGBoost with and without odds confirms that the market odds themselves are the most informative feature.

4.4 Probability calibration

In betting applications, calibration of predicted probabilities matters as much as ranking or accuracy. We therefore calibrate the XGBoost model with odds using isotonic regression via `CalibratedClassifierCV` (3-fold CV on train+validation).

On the test set, calibration improves log-loss from ≈ 1.0319 to ≈ 1.0156, while accuracy remains around 0.49.

Table 7: XGBoost performance before and after calibration.

model	$\log_l oss_t est$	accuracy$_t est$
XGB with odds (uncalibrated)	1.0319	0.4976
XGB with odds (isotonic calibrated)	1.0156	0.4901

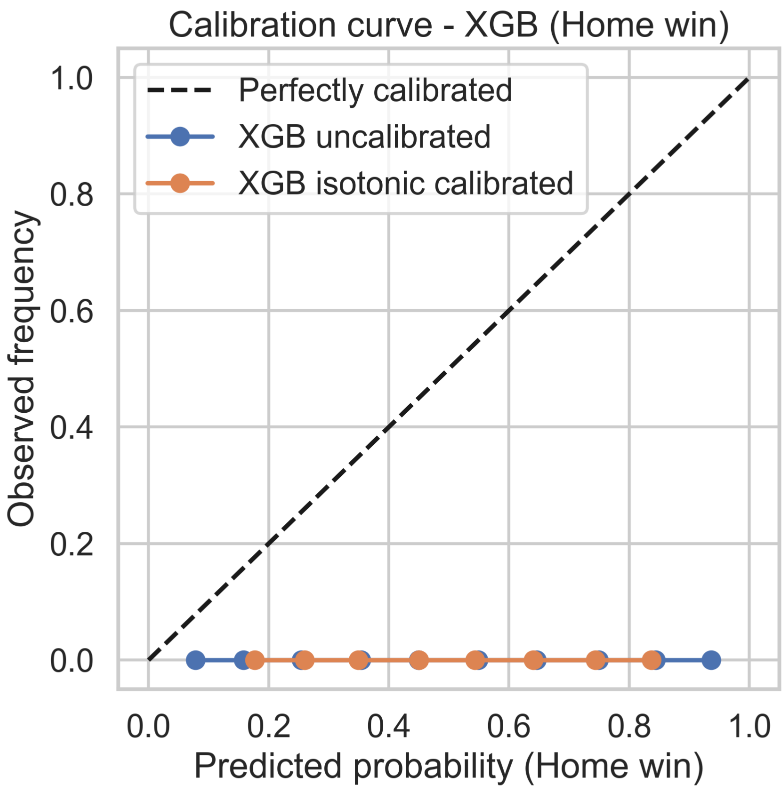


Figure 8: Calibration curves for XGBoost (home win) before and after isotonic calibration.

5 From Probabilities to Bets

5.1 Embedding bets and expected value

For each match in the test set and each outcome $o \in \{H, D, A\}$, we form a candidate bet with:

- model probability $p_{o,t}$ (from logistic regression or calibrated XGBoost),
- bookmaker odds $c_{o,t}$ (Bet365),
- expected value $\mathrm{EV}_{o,t} = p_{o,t} \cdot c_{o,t} - 1$.

This yields a `bets` table with one row per match and outcome.

Table 8: Descriptive statistics of EV (logistic regression model).

stat	value
count	6837.0
mean	-0.0581
std	0.1624
min	-0.8122
max	0.7558

Table 9: Descriptive statistics of EV (calibrated XGBoost model).

stat	value
count	6837.0
mean	0.0375
std	0.297
min	-0.6396
max	3.8806

For logistic regression, the mean EV across all candidate bets is negative (around -0.058). For XGBoost after calibration, the mean EV becomes slightly positive (around $+0.037$), suggesting that the model "believes" many bets to be profitable on average.

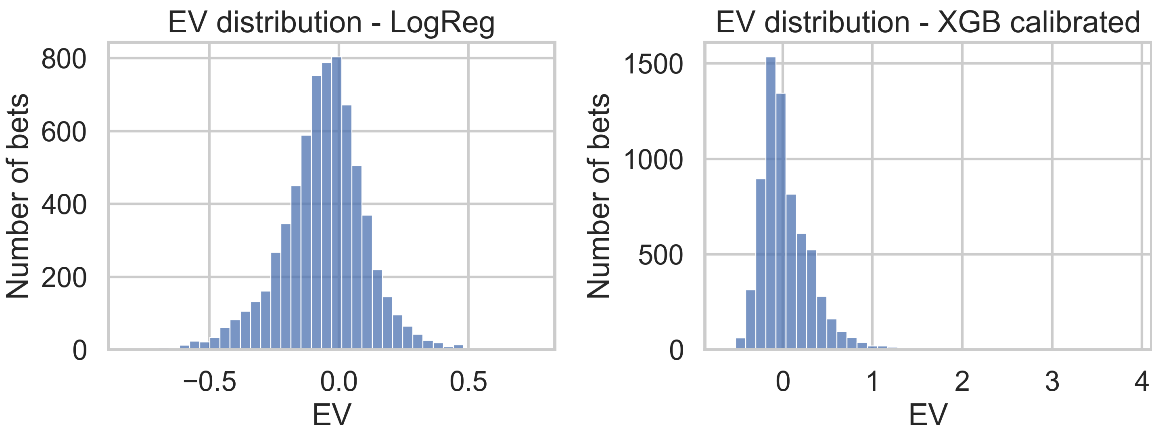


Figure 9: Distributions of expected value (EV) for logistic regression and calibrated XGBoost.

5.2 Team-pair probability matrices

For additional insight, we aggregate probabilities by home-away team pairs (i, j) to build matrices:

- M_{ij}^{market}: average market probability of home win (Bet365 implied),
- M_{ij}^{model}: average model probability of home win.

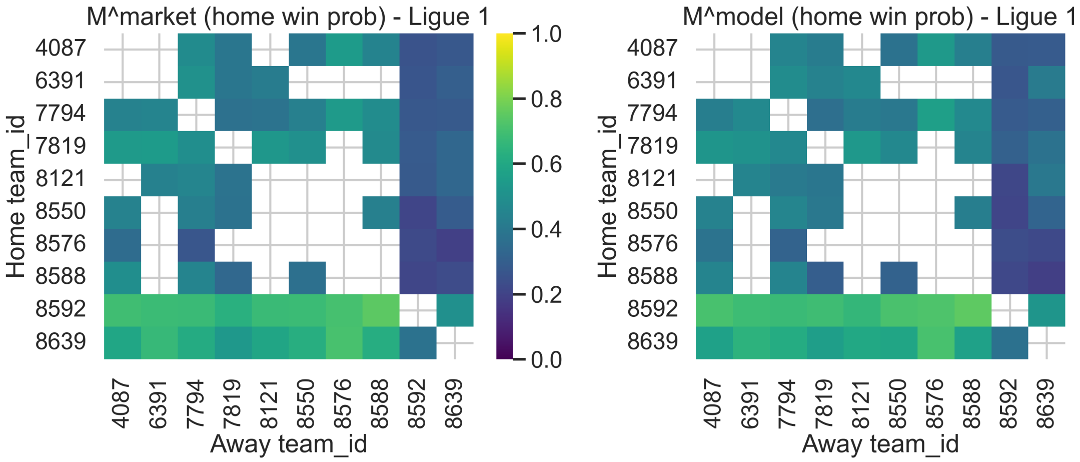


Figure 10: Home win probability matrices for Ligue 1: market vs model (subset of teams).

These matrices allow us to identify persistent disagreements between model and market, although, as shown later, such disagreements do not translate into profitable strategies.

6 Betting Strategies and Optimisation

6.1 Simple value betting rules

We first consider simple, transparent value betting rules on the test set, with unit stake per bet and initial bankroll of 100 units:

R1 Bet on all outcomes with EV > 0.

R2 Bet on all outcomes with EV > 0.05 (at least 5% edge).

R3 For each match, consider only the favourite according to odds (lowest odds) and bet if EV > 0.

Table 10: Backtest results for value betting rules using logistic regression.

strategy	n$_{b}ets$	final$_{b}ankroll$	ROI	max$_{d}rawdown$	avg$_{g}ain_{p}er_{b}et$
LogReg R1 (EV>0)	2407	-85.62	-1.8562	-1.9185	-0.0771
LogReg R2 (EV>0.05)	1525	5.19	-0.9481	-1.0322	-0.0622

All strategies produce highly negative ROIs. For example, R1 (EV>0) makes 2407 bets and ends with a bankroll around -85.6 units (ROI $\approx -186\%$), with a maximum drawdown close to -192%.

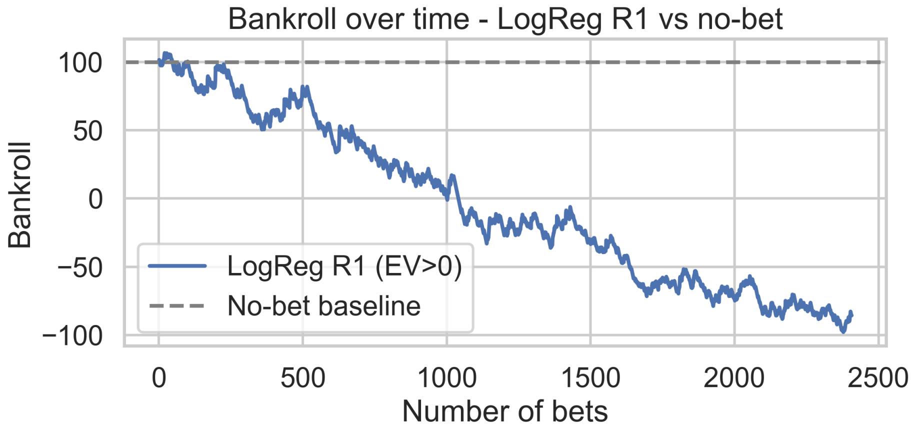


Figure 11: Bankroll evolution for logistic regression R1 strategy vs no-bet baseline.

6.2 Strategies with calibrated XGBoost

We repeat the same value betting rules with calibrated XGBoost probabilities, and add more conservative filters:

- EV threshold (EV > 0.2),
- odds range restriction (e.g. odds between 1.5 and 5.0),
- confidence filter based on disagreement between model and market $|p_{\text{model}} - q_{\text{market}}|$.

Table 11: Backtest results for strategies using calibrated XGBoost probabilities.

strategy	$\mathrm{n}_b ets$	$\mathrm{final}_b ankroll$	ROI	$\mathrm{max}_d rawdown$	$\mathrm{avg}_g ain_p er_b et$
XGB calib R1 (EV>0)	3054	-78.9	-1.789	-1.6553	-0.0586
XGB calib R2 (EV>0.05)	2575	-55.52	-1.5552	-1.5187	-0.0604
XGB calib ultra-conservative	955	37.93	-0.6207	-0.8542	-0.065
XGB calib EV+confidence	551	61.23	-0.3877	-0.6334	-0.0704

Even with strict thresholds and confidence filters, all strategies remain loss-making. For instance:

- R1 (EV>0): 3054 bets, ROI $\approx -179\%$.
- R2 (EV>0.05): 2575 bets, ROI $\approx -156\%$.
- Ultra-conservative strategy (EV>0.2, odds between 1.5 and 5.0, two leagues): 955 bets, ROI $\approx -62\%$.
- EV+confidence strategy (EV>0.1, $|p_{\text{model}} - q_{\text{market}}| > 0.1$, odds between 1.5 and 5.0): 551 bets, ROI $\approx -39\%$.

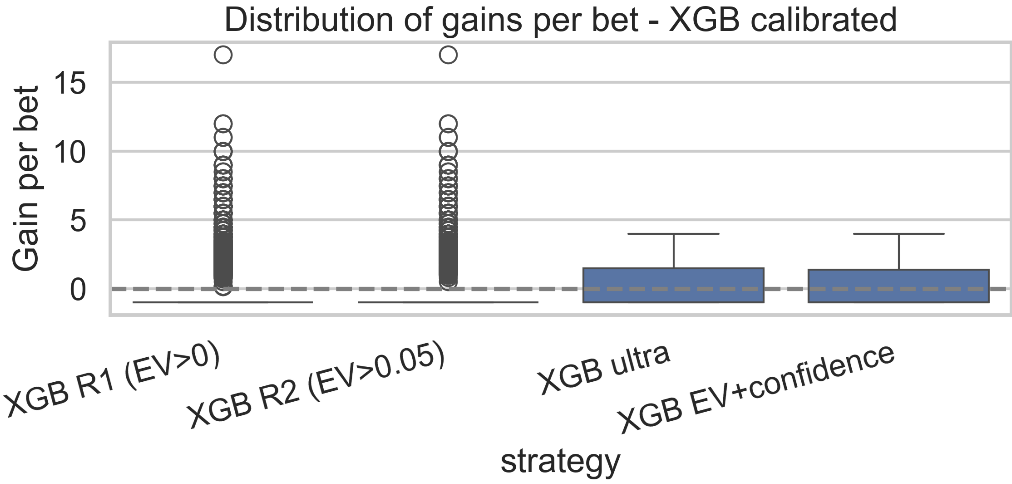


Figure 12: Distribution of gain per bet for selected strategies using calibrated XGBoost.

6.3 Convex portfolio optimisation

To connect with convex optimisation ideas, we consider the following static portfolio problem over a universe of candidate value bets (e.g. EV>0.05 in Ligue 1):

$$\begin{aligned} \max_{w_1,\ldots,w_n} \quad & \sum_{i=1}^{n} w_i \cdot \mathrm{EV}_i \\ \text{s.t.} \quad & \sum_{i=1}^{n} w_i \leq B, \\ & 0 \leq w_i \leq w_{\max} \quad \forall i, \end{aligned}$$

where B is the total budget (100 units) and $w_{\max}$ caps exposure per bet (e.g. 5 units). This is a linear program, solved with `scipy.optimize.linprog`.

On the Ligue 1 universe with 607 candidate bets, the optimiser allocates the full budget $B = 100$ and, under model-estimated EVs, achieves a theoretical expected gain of about 37 units. However, when these allocations are backtested using actual match outcomes, realised returns remain negative, confirming that EVs are overestimated by the model.

7 Discussion and Business Implications

7.1 Why the model is not profitable

Despite reasonable predictive performance (log-loss close to 1.0 and accuracy around 50%), none of the tested betting strategies produces a positive ROI on historical data. Several factors explain this gap between statistical and financial performance:

- **Market efficiency**: Bet365 implied probabilities already provide a very strong baseline. XGBoost without odds performs significantly worse, confirming that most predictive signal resides in the odds themselves.

- **Limited data**: The dataset covers only two leagues and one bookmaker over a finite period, with no information on closing odds, line movement, injury news, or advanced performance metrics (e.g. expected goals).
- **High variance**: Football outcomes are inherently noisy; even small systematic biases in probabilities are hard to exploit reliably without a very large edge.
- **Calibration limitations**: Isotonic calibration improves log-loss but does not fully correct systematic overestimation of EVs on high-odds, high-variance bets.

7.2 Business value of the approach

Even though no profitable strategy was found, the project delivers several business-relevant insights:

- A reusable **end-to-end pipeline** from raw match data and odds to probabilistic forecasts, expected values, and simulated ROIs.
- A concrete demonstration that **good ML metrics do not guarantee economic profitability**. This is crucial when deploying ML systems in finance or operations.
- A methodological template for evaluating new data sources: any additional features (e.g. proprietary tracking data) could be plugged into this pipeline and assessed directly in terms of ROI.

Given the negative backtest results, a rational business decision would be to *not* deploy this strategy with real money, unless significantly richer data and models are available.

Moreover, the backtests abstract away from practical frictions such as stake limits, account restrictions, and transaction costs, which would further reduce the economic viability of any strategy.

7.3 Limitations and future work

Key limitations include:

- reliance on historical, static data and a single bookmaker,
- absence of transaction costs, bet size constraints, and account limitations,
- no modelling of time-varying market dynamics (closing odds, in-play information).

Future work could explore:

- integrating richer features (expected goals, player-level metrics, injuries, scheduling effects),
- incorporating market microstructure (line movement, multiple bookmakers, closing line value),
- using hierarchical or Bayesian models to better capture uncertainty and reduce overfitting.

8 Conclusion

This project set out to "beat the bookmakers" using machine learning on football match data and Bet365 odds. We constructed a comprehensive dataset for Premier League and Ligue 1, engineered form and team attribute features, trained probabilistic classifiers (logistic regression, XGBoost), calibrated their probabilities, and transformed predictions into betting strategies and portfolio allocations.

Empirically, the best model (calibrated XGBoost with odds) achieved solid probabilistic performance, but all tested betting strategies – from naive value betting to ultra-conservative filters and linear portfolio optimisation – produced substantially negative ROIs on historical data. This provides strong evidence that, given the available data, the Bet365 market is too efficient to be beaten by this class of models.

The main contribution is therefore not a profitable strategy, but a rigorous, data-driven demonstration of *why* such a strategy is elusive with public historical data, and a general methodology for assessing ML-driven trading or betting ideas in a business context.

This project illustrates not only technical modelling but also business judgment, by explicitly linking model performance to ROI and deployment decisions.