Whistles or Whispers - Betting Market Anomaly Study Final Report

Eren Batu Cansever
January 2026
batu.cansever@sabanciuniv.edu

I. Motivation and Intro

Football betting markets are among the largest and most liquid markets in the world, with billions of dollars wagered every day across leagues, bookmakers, and exchanges. Acording to 2025 estimates, only in Turkey, approximately 60 million USD is wagered daily through legal bookmakers, supported by a base of over 15 million active bettors. These figures alone highlight the extraordinary scale, liquidity, and influence of betting markets to people. The motivation of this research is to observe and analyze potential suspicious patterns in such environments. Specifically, the study examines whether heavily bet, bookmaker-supported favorites shows unexpected underperformance relative to their implied probabilities, and whether abnormal line movements in matches with extreme betting concentration. The goal is not to make allegations, but to identify, measure, and document anomalies that may indicate hidden risk, bias, or structural inefficiencies within otherwise highly liquid football betting markets. All code, and experiment notebooks used in this study are available in the project repository [1].

II. Datasets

This study is based on two complementary football betting datasets, each serving an important role in the analysis. The first dataset was obtained from football-data.co.uk via Kaggle. . It contains opening odds, match scores, and final outcomes for all seasons of the five major European leagues spanning from 2005/2006 to 2018/2019. Due to its long historical coverage and consistent structure, this dataset is used primarily to analyze opening odds and their implied probabilities across leagues and seasons. The kaggle link is: https://www.kaggle.com/datasets/jamieandrews/footballdatacouk

The second dataset is the Beat The Bookie dataset, which includes a richer set of betting-related variables such as opening odds, closing odds, match scores, and results. While the dataset provides extensive information, this study focuses exclusively on closing odds. These closing prices are used to compute line movement between opening and closing markets, enabling the identification of matches with extreme betting concentration and helping to proxy the most heavily bet sides. The kaggle link fort his dataset is: https://www.kaggle.com/datasets/austro/beat-the-bookie-worldwide-football-dataset

III. Data Preprocessing and Exploratory Preparation

A. Data Sources and Scope

Two complementary data sources are employed in this study:

- **Football-Data (FBD):** Match results and opening odds collected from multiple bookmakers.
- **Beat-The-Bookie (BTB):** Aggregated closing odds obtained from a broad set of bookmakers.

The analysis is restricted to the *Big Five* European leagues:

- England – Premier League
- Germany – Bundesliga
- Italy – Serie A
- France – Ligue 1
- Spain – Primera Division

The temporal coverage is defined as follows:

- **FBD:** Group1 (2009–2015) later used for line movement calculations, Group2 (2013–2019) used for oppening odds only EDA.
- **BTB:** (2009–2015) aligned with Group1.

B. Football-Data (FBD): Opening Odds

1) File Discovery and Loading: All CSV files contained in the Football-Data repository are discovered and loaded. Season information is extracted from file names, and only files corresponding to the predefined GROUP1 and GROUP2 periods are retained. These files are combined into one dataset.

2) Construction of Average Opening Odds: For each match, Football-Data provides opening odds from multiple bookmakers for the home win (H), draw (D), and away win (A) outcomes. To obtain a representative market price, odds are averaged across bookmakers for each outcome. Missing values are excluded from the averaging procedure, and observations with incomplete odds information are removed.

3) Standardization and Cleaning: Several normalization steps are applied to ensure consistency and reliable matching:

- **League mapping:** Division codes are mapped to full league names (e.g., E0 $\rightarrow$ England: Premier League), and only Big Five leagues are retained.
- **Team names:** Team name variants and special cases are fitted.
- **Match dates:** Dates are parsed using multiple possible formats and converted to the standard YYYY-MM-DD format. Records with unparseable dates are excluded.

- **Odds formatting:** All odds are rounded to two decimal places.

4) Output Dataset: The resulting opening odds dataset is stored to use on line movement calcutaions and feeding the ML model. File name is:

`average_opening_odds_group1.csv`

These are the states of the dataset after the mentioned steps are applied as shown in Fig. 1 and Fig. 2.

	HomeTeam	AwayTeam	FTR	AVG_H	AVG_D	AVG_A	Season
552	Barnsley	Crawley Town	A	2.05	3.33	3.57	2014-2015
553	Bradford	Coventry	H	2.37	3.29	2.95	2014-2015
554	Colchester	Oldham	D	2.59	3.21	2.69	2014-2015
555	Fleetwood Town	Crewe	H	1.94	3.39	3.89	2014-2015
556	Leyton Orient	Chesterfield	A	1.96	3.37	3.81	2014-2015

Fig. 1. Dataset state for opening odds.

	HomeTeam	AwayTeam	Date	league	AVG_H	AVG_D	AVG_A
0	Accrington	Aldershot	2010-08-07	England: League Two	2.44	3.24	2.65
1	Burton	Oxford	2010-08-07	England: League Two	2.28	3.26	2.88
2	Bury	Port Vale	2010-08-07	England: League Two	2.21	3.24	3.01
3	Chesterfield	Barnet	2010-08-07	England: League Two	1.64	3.51	5.09
4	Crewe	Hereford	2010-08-07	England: League Two	1.87	3.36	3.79

Fig. 2. Dataset state for opening odds.

C. Beat-The-Bookie (BTB): Closing Odds

1) Loading and Filtering: The Beat-The-Bookie dataset (`closing_odds.csv.gz`) is loaded and filtered to match the analytical scope. Only observations within the 2009-08-01 to 2015-06-30 period and belonging to the Big Five leagues are retained.

2) Variable Selection: Only the variables required for matching and analysis are selected. Column names are standardized to align with the FBD (openning odds) schema. Match dates are converted to datetime format, and observations with missing closing odds are removed.

3) Output Dataset: The cleaned closing odds dataset is stored as:

`average_closing_odds_group1.csv`

D. Matching Opening and Closing Odds

1) Join Keys and Coverage: Opening and closing odds datasets are merged using an inner join on the following keys:

- `HomeTeam`
- `AwayTeam`
- `Date`
- `league`

After merge, checked the merged dataset to evaluate matching coverage and catch potential key mismatches.

2) Definition of Line Movement: Line movement (LM) is defined as the difference between closing and opening odds:

$$\begin{aligned} LM_{\text{Home}} &= AVG\_H_{\text{Close}} - AVG\_H, \\ LM_{\text{Draw}} &= AVG\_D_{\text{Close}} - AVG\_D, \\ LM_{\text{Away}} &= AVG\_A_{\text{Close}} - AVG\_A. \end{aligned} \tag{1}$$

A positive value indicates that the closing odds increased relative to the opening odds, implying that the corresponding outcome was priced as less likely by the market at closing.

Example of a line movement cell shown in Fig. 3.

LM_Home	LM_Draw	LM_Away
-0.0997	-0.0883	0.0714
0.1524	0.0228	-0.2248
-0.1962	0.0223	0.1438

Fig. 3. Calculated line movements of teams.

3) League-Level Summary: Average line movements are computed at the league level to facilitate cross league comparison of market dynamics.

E. Assumptions and Design Choices

The preprocessing pipeline is guided by the following design decisions: the bookmaker average is used as a representative of the market's common thought; team name and date normalization are treated as mandatory to ensure accurate matching; the analysis is restricted to the Big Five leagues to preserve data quality and comparability; and line movement is defined as closing minus opening odds, with probability conversion and overround adjustments deferred to subsequent modeling stages.

Final states of the dataset before the EDA shown in Fig. 4 and Fig. 7:

	HomeTeam	AwayTeam	Date	AVG_H_Close	AVG_D_Close	AVG_A_Close	league
122419	Wolfsburg	VfB Stuttgart	2009-08-07	1.8977	3.4923	3.9031	Germany: Bundesliga
122643	Auxerre	Sochaux	2009-08-08	2.0496	3.0676	3.8108	France: Ligue 1
122644	Grenoble	Marseille	2009-08-08	5.0077	3.3435	1.7277	France: Ligue 1

Fig. 4. Teams matched with their closing odds.

IV. Exploratory Data Analysis (EDA)

A. Data Health Checks

1) Data Health Checks: Basic data health checks are performed prior to analysis:

	HomeTeam	AwayTeam	FavSide	FavTeam	pFav_win	FavOdds	FTR	FavCorrect	MaxAbsLM
3	Bochum	Bayern Munich	A	Bayern Munich	0.668490	1.4004	A	1	0.8272
8	Juventus	Inter	H	Juventus	0.689990	1.3693	H	1	1.0236
10	Juventus	Atalanta	H	Juventus	0.718586	1.3152	H	1	7.7193
16	Sunderland	Manchester City	A	Manchester City	0.590062	1.6007	H	0	0.4993
24	Arsenal	Norwich	H	Arsenal	0.788500	1.1967	D	0	0.6367

Fig. 5. Calculated line movements of teams with Favorite Team variable.

- File and season discovery, season-level distributions, and row–column dimensions are reviewed.
- Missing value checks are conducted; observations with missing odds values are removed.
- Potential duplicates in merge keys are identified and documented prior to dataset integration.

B. From Odds to Implied Probabilities and Overround

Opening odds are converted into implied probabilities according to:

$$p^{raw} = \frac{1}{\text{odds}}$$

The bookmaker overround (margin) is defined as:

$$\text{Overround} = p_H^{raw} + p_D^{raw} + p_A^{raw}$$

To remove the margin effect, normalized probabilities are computed as:

$$p_i = \frac{p_i^{raw}}{p_H^{raw} + p_D^{raw} + p_A^{raw}}, \quad i \in \{H, D, A\}$$

1) Validation Checks: The distribution of overround values is examined, and numerical checks confirm that normalized probabilities satisfy:

$$p_H + p_D + p_A \approx 1$$

C. Favorite Definition and Overall Performance

The favorite side is defined by comparing implied probabilities for home and away wins. For each match, the following variables are derived:

- `FavOdds_fixed`: odds of the favorite
- `pFav_win`: implied probability of favorite win
- `FavCorrect`: indicator of whether the favorite won the match

Overall favorite performance is evaluated by comparing:

- The realized win rate of favorites (Actual)
- The average implied win probability (Expected)

D. Favorite Strength Categories

Favorites are further categorized based on their odds into the following strength groups:

- Extremely Heavy: < 1.30
- Heavy: 1.30–1.40
- Strong: 1.40–1.60
- Moderate: 1.60–1.70
- Weak: > 1.70

For each category, the number of matches, actual win rate, expected win rate, and their difference are computed.

E. Seasonal Breakdown, Residuals, and Calibration

1) Seasonal Performance: At the season level, actual and expected favorite win rates are computed and compared. Additional seasonal diagnostics include average favorite odds and mean overround.

2) Residual Analysis: Residuals are defined as:

$$\text{Residual} = \text{Actual\_Win} - p_{\text{Fav\_win}}$$

The distribution of residuals is summarized using descriptive statistics such as the median and standard deviation.

3) Calibration Analysis: To assess probability calibration, matches are grouped into bins of width 0.05 based on their expected favorite win probability. For each bin, the following are reported:

- Number of matches
- Average expected win probability
- Observed win rate
- Difference between observed and expected rates

F. Line Movement Exploratory Analysis

1) Definition of Line Movement: As explained before opening odds from FBD and closing odds from BTB are combined to define line movement:

$$LM_{\text{Home}} = AVG\_H_{\text{Close}} - AVG\_H,$$
$$LM_{\text{Draw}} = AVG\_D_{\text{Close}} - AVG\_D,$$
$$LM_{\text{Away}} = AVG\_A_{\text{Close}} - AVG\_A$$

Positive values indicate an increase in odds at closing, while negative values indicate odds shortening.

G. Outlier Analysis: Heavy Favorite Non-Wins

For heavy favorites, the rate of non-wins (draws or losses) is compared against the expected non-win probability. The most extreme cases—matches with high expected favorite win probabilities that nevertheless resulted in non-wins—are identified. The top 20 such cases are documented, including season, teams, favorite side, odds, implied probabilities, and final result.

V. Hypothesis Testing

This section examines whether observed match outcomes for selected subsamples are statistically consistent with the implied win probabilities derived from closing odds. The analysis focuses on heavy favorites and matches with high line movement, under a probabilistic framework based on match-specific implied probabilities.

A. Test Design and Sample Definition

1) Sample Groups: The following subsamples are considered:

- **Heavy favorites:** matches with `FavOdds_fixed` $<$ 1.40.
- **High line movement:** matches belonging to the upper 30% of the line-movement distribution, where line movement is defined according to the selected metric. Note that high line movement means high betting volume.

Implied probabilities are obtained from **closing odds (BTB)** and adjusted for bookmaker margin (overround). The implied probability of a favorite win for match i is denoted by $p_i = p_{\text{Fav\_win},i}$.

	HomeTeam	AwayTeam	FTR	FavTeam	FavCorrect
552	Barnsley	Crawley Town	A	Barnsley	0
553	Bradford	Coventry	H	Bradford	1
554	Colchester	Oldham	D	Colchester	0
555	Fleetwood Town	Crewe	H	Fleetwood Town	1
556	Leyton Orient	Chesterfield	A	Leyton Orient	0
557	Milton Keynes Dons	Gillingham	H	Milton Keynes Dons	1
558	Port Vale	Walsall	D	Port Vale	0
559	Preston	Notts County	D	Preston	0
560	Rochdale	Peterboro	A	Peterboro	1
561	Sheffield United	Bristol City	A	Sheffield United	0

Fig. 6. Heavy favorites subgroup.

	HomeTeam	AwayTeam	Date	league	LM_Home	LM_Draw	LM_Away	MaxAbsLM	FTR
10	Juventus	Atalanta	2014-05-05	Italy: Serie A	0.1452	-1.7289	-7.7193	7.7193	H
203	Dortmund	Freiburg	2012-05-05	Germany: Bundesliga	-0.0904	1.1756	2.3796	2.3796	H
200	Dortmund	Mainz	2012-03-03	Germany: Bundesliga	0.0354	-0.2396	-1.5539	1.5539	H
54	Marseille	Nice	2012-11-11	France: Ligue 1	0.0804	-0.2768	-1.3711	1.3711	D
95	Marseille	Boulogne	2009-12-12	France: Ligue 1	-0.0133	0.2946	1.1250	1.1250	H
81	Inter	Parma	2015-04-04	Italy: Serie A	-0.0017	-0.0969	-1.1159	1.1159	D
204	Hannover	Kaiserslautern	2012-05-05	Germany: Bundesliga	0.0526	-0.2270	-1.0707	1.0707	H
43	Lazio	Bologna	2013-05-05	Italy: Serie A	-0.1093	0.2521	1.0643	1.0643	H
173	Almeria	Real Madrid	2014-12-12	Spain: Primera Division	-1.0525	0.0104	-0.0125	1.0525	A
73	Hoffenheim	Bayern Munich	2013-03-03	Germany: Bundesliga	-1.0521	0.1214	-0.0043	1.0521	A

Fig. 7. High Line movement subgroup.

2) Model Assumption:

3) Model Assumption: For each match i, the favorite-win outcome $Y_i \in \{0, 1\}$ is modeled as:

$$Y_i \sim \text{Bernoulli}(p_i),$$

where the success probability p_i varies across matches. Conditional independence across matches is assumed given the set of probabilities $\{p_i\}$.

4) Robustness and Data Quality Controls: Several robustness measures are applied to hypothesis testing:

- Implied probabilities are clipped to the interval $[10^{-6}, 1-10^{-6}]$ to avoid numerical instability.
- Observations with missing odds values are excluded from the analysis.

B. Monte Carlo Simulation Test

A simulation-based hypothesis test is conducted for the heavy-favorite subsample and, separately, for the high line-movement subsample.

1) Hypotheses: Let $W = \sum_{i=1}^{n} Y_i$ denote the total number of favorite wins in a given subsample.

- H_0: observed outcomes are consistent with the implied probabilities $\{p_i\}$.
- H_1: observed outcomes deviate significantly from expectations, indicating over- or under-performance.

2) Test Statistic: The test statistic is defined as the observed number of favorite wins:

$$W_{\text{obs}} = \sum_{i=1}^{n} Y_i.$$

3) Simulation Procedure: The null distribution of W is generated via Monte Carlo simulation:

1) The number of simulations is set to $N_{\text{SIMS}} = 20{,}000$.
2) For each simulation s, outcomes are generated independently as $Y_i^{(s)} \sim \text{Bernoulli}(p_i)$.
3) The simulated win count is computed as $W^{(s)} = \sum_{i=1}^{n} Y_i^{(s)}$.

The expected value and variance of W under the null hypothesis are given by:

$$E[W] = \sum_{i=1}^{n} p_i, \quad \text{Var}(W) = \sum_{i=1}^{n} p_i(1-p_i),$$

with $\text{Std}(W) = \sqrt{\text{Var}(W)}$.

4) P-values and Decision Rule: One-sided and two-sided p-values are computed as:

$$p_{\text{left}} = P(W^{(s)} \le W_{\text{obs}}), \quad p_{\text{right}} = P(W^{(s)} \ge W_{\text{obs}}),$$

$$p_{\text{two}} = 2 \cdot \min\left(p_{\text{left}}, p_{\text{right}}\right).$$

The significance level is set to $\alpha = 0.05$. The decision rules are:

- $p_{\text{left}} < \alpha$: evidence of underperformance.
- $p_{\text{right}} < \alpha$: evidence of overperformance.
- Otherwise, H_0 is not rejected.

a) Summary of Findings.: Based on the notebook outputs, the null hypothesis is not rejected for either the heavy-favorite or high line-movement subsample ($p \ge 0.05$).

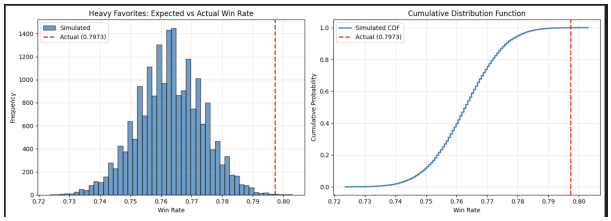

Fig. 8. Heavy favorites subgroup Monte Carlo.

5) Model Assumption:

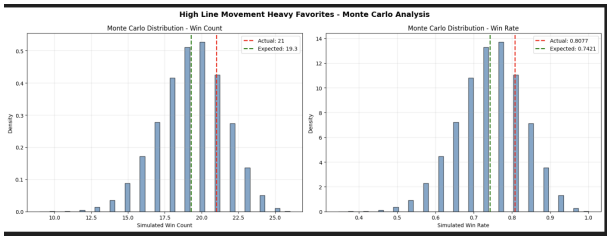

Fig. 9. High Movement subgroup Monte Carlo.

C. Goodness-of-Fit Test

1) Objective and Hypotheses: A binned goodness-of-fit test is applied to assess consistency between expected and observed favorite wins.

- H_0: expected and observed win counts across probability bins follow the same distribution.
- H_1: a statistically significant discrepancy exists.

2) Procedure: Matches are grouped into probability bins (e.g., from $[0.35, 1.00)$ with a bin width of 0.05). For each bin b:

$$E_b = \sum_{i \in b} p_i, \quad O_b = \sum_{i \in b} Y_i.$$

A chi-square goodness-of-fit statistic is computed as:

$$\chi^2 = \sum_b \frac{(O_b - E_b)^2}{E_b},$$

subject to standard minimum expected-count requirements. Adjacent bins are merged where necessary. The test is conducted at $\alpha = 0.05$.

a) Summary of Findings.: As reported in the project documentation, the null hypothesis is not rejected ($p \geq 0.05$). Final test statistics and p-values should be reported in Table **??**.

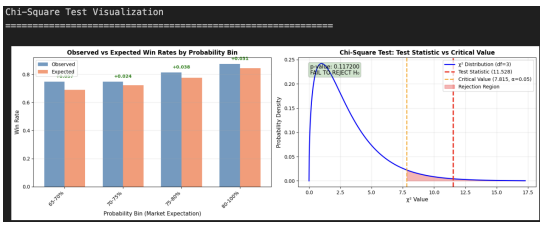

Fig. 10. Chi-Square Visualization.

D. Kolmogorov–Smirnov Test

1) Objective and Hypotheses: The Kolmogorov–Smirnov (KS) test evaluates whether the distribution of observed outcomes in the high line-movement subsample is consistent with the theoretical distribution implied by the probabilities $\{p_i\}$.

- H_0: observed and theoretical distributions are consistent.
- H_1: a statistically significant difference exists.

2) Procedure: A theoretical sample is generated by simulating Bernoulli outcomes using the probabilities $\{p_i\}$ over multiple repetitions (e.g., 1,000 draws). A two-sample KS test is then applied. Complementary summaries based on probability bins are also reported.

The significance level is $\alpha = 0.05$. Effect size interpretation follows:

$$D < 0.1 \text{ (negligible)}, \quad D < 0.3 \text{ (small)},$$
$$D < 0.5 \text{ (medium)}, \quad D \geq 0.5 \text{ (large)}.$$

a) Summary of Findings.: According to the reported outputs, the null hypothesis is not rejected ($p \geq 0.05$), with effect sizes in the negligible-to-small range.

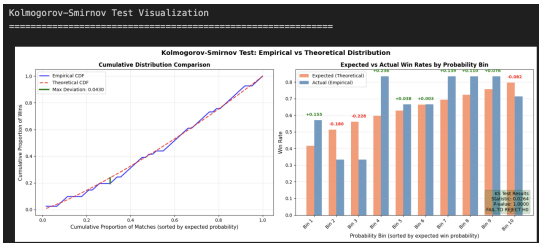

Fig. 11. Kolmogorov-Smirnov Test Visualization.

E. Sensitivity and Robustness Analysis

Several sensitivity analyses may be conducted to assess robustness:

- **Threshold sensitivity:** alternative heavy-favorite thresholds (1.30, 1.35, 1.45) and line-movement percentiles (60th, 70th, 80th).
- **Temporal and league splits:** re-estimation by season and by league.
- **Opening vs. closing probabilities:** comparison with probabilities derived from opening odds.
- **Data quality effects:** evaluation of merge coverage and duplicate filtering.

VI. Machine Learning Model

This section shows the end-to-end baseline machine learning pipeline used to predict outcomes for heavy favorites The pipeline mirrors the implementation in `src/baseline_upset_model.py` and formalizes data contracts, feature construction, evaluation protocol, and reproducibility considerations.

A. Objective and Prediction Target

The prediction target is defined as:

$$\texttt{HeavySide\_Won} \in \{0, 1\},$$

indicating whether the market-defined favorite (home or away, with draws excluded from favorite selection) wins the match.

The modeling objective is to assess whether signals embedded in betting odds—particularly for strong favorites—translate into realized match outcomes and to evaluate the reliability of these signals under a probabilistic classification framework.

B. Data Contracts

1) Input Sources: The pipeline consumes two preprocessed datasets:

- **Opening odds:** `processed/average_opening_odds_group1.csv` (Football-Data)
- **Closing odds:** `processed/average_closing_odds_group1.csv` (Beat-The-Bookie)

2) Required Columns: The following columns are required:

- **Opening odds:** `HomeTeam`, `AwayTeam`, `Date`, `league`, `FTR`, `AVG_H`, `AVG_D`, `AVG_A`
- **Closing odds:** `HomeTeam`, `AwayTeam`, `Date`, `league`, `AVG_H_Close`, `AVG_D_Close`, `AVG_A_Close`

Datasets are merged using an inner join on the composite key:

$$(\texttt{HomeTeam}, \texttt{AwayTeam}, \texttt{Date}, \texttt{league}).$$

3) Cleaning Rules: The following data quality rules are enforced:

- Observations with missing odds or missing match results are removed.
- Odds are required to be strictly positive and greater than 1.001.
- Date formats and team names are assumed to be normalized upstream during preprocessing and EDA.

C. Feature Engineering

1) Implied Probabilities and Overround: Opening odds are converted into implied probabilities as follows:

$$p_H^{raw} = \frac{1}{AVG\_H}, \quad p_D^{raw} = \frac{1}{AVG\_D}, \quad p_A^{raw} = \frac{1}{AVG\_A}.$$

The opening overround is defined as:

$$\text{Overround}_{open} = p_H^{raw} + p_D^{raw} + p_A^{raw} - 1.$$

Vig-free probabilities are obtained by normalization:

$$p_i = \frac{p_i^{raw}}{p_H^{raw} + p_D^{raw} + p_A^{raw}}, \quad i \in \{H, D, A\}.$$

The same procedure is applied to closing odds to obtain p_H^{close}, p_D^{close}, p_A^{close}, and Overround_{close}.

2) Favorite Definition and Probability Shift: The favorite side is defined by comparing implied win probabilities (draw excluded):

$$\texttt{FavSide} = \begin{cases} H, & \text{if } p_H \geq p_A, \\ A, & \text{otherwise.} \end{cases}$$

Favorite win probabilities are defined as:

$$p_{Fav}^{open} = \begin{cases} p_H, & \text{if FavSide = H,} \\ p_A, & \text{if FavSide = A,} \end{cases} \qquad p_{Fav}^{close} = \begin{cases} p_H^{close}, \\ ifFavSide = \\ H, \\ p_A^{close}, \\ ifFavSide = \\ A. \end{cases}$$

The probability shift between opening and closing is captured by:

$$\Delta p_{Fav} = p_{Fav}^{close} - p_{Fav}^{open}.$$

3) League Context: Categorical league information is encoded using one-hot dummy variables derived from `league`, with the first category dropped to avoid multicollinearity.

D. Heavy Favorite Filter and Label Construction

1) Heavy Favorite Definition: A heavy favorite is defined using the opening odds threshold:

$$\texttt{FavOdds\_fixed} < 1.40.$$

2) Label Definition: The binary label is constructed as:

$$\texttt{HeavySide\_Won} = \begin{cases} 1, if FavSide = H and FTR = H, or \\ FavSide = A and FTR = A, \\ 0, otherwise. \end{cases}$$

Draws are excluded from favorite selection but are treated as non-win outcomes in evaluation.

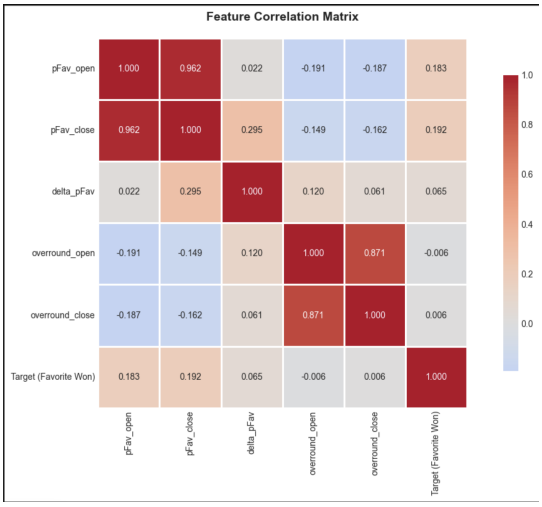

Fig. 12. Favorites Correlation.

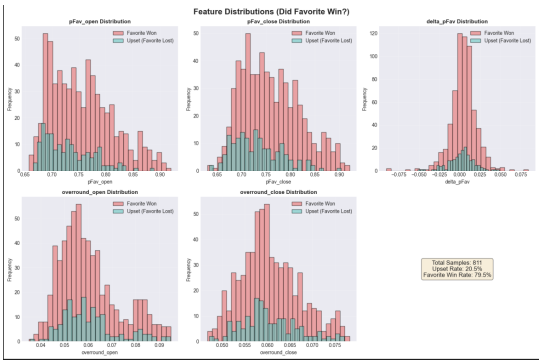

Fig. 13. Favorite Distrubution.

E. Train–Test Protocol

The heavy-favorite subset is split into training and test sets using a stratified 75/25 split with a fixed random seed (42). Class balance is preserved across splits.

An optional maximum sample size (e.g., 20,000 observations) may be applied prior to splitting to support faster experimentation.

Closing-odds features are included by design, corresponding to a prediction scenario at or near kickoff. Time-aware splits are deferred to future extensions.

F. Model Specification and Evaluation

1) Model: The baseline classifier is a logistic regression model with:

- Solver: `liblinear`
- Regularization: L2
- Class weighting: `balanced`

The model outputs predicted probabilities for the positive class (`HeavySide_Won = 1`).

2) Evaluation Metrics: Model performance is evaluated using:

- Area Under the ROC Curve (AUC)
- Brier score
- Accuracy
- Positive class rate

For interpretability, model coefficients are reported to assess the direction and magnitude of feature contributions.

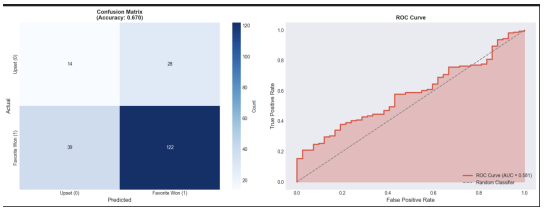

Fig. 14. Roc Curve and Confusion Matrix.

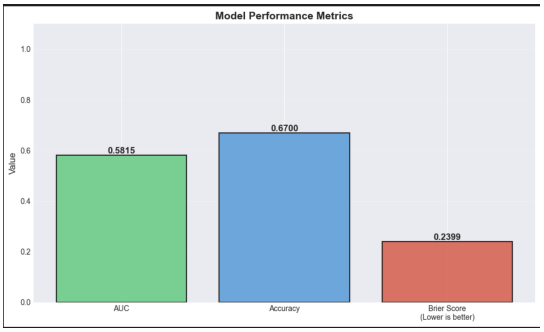

Fig. 15. Brier Score, AUC, Accuracy .

G. Edge Cases and Data Filters

The pipeline explicitly handles several edge cases and data quality filters. Invalid odds values (≤ 1.001) are removed; observations with missing closing odds are dropped prior to merging; duplicate join keys are checked upstream and resolved via de-duplication or aggregation; and extreme overround values are monitored and may be capped or excluded in robustness checks.

H. Reproducibility and Configuration

Pipeline configuration is controlled via environment variables:

- `OPENING_CSV`, `CLOSING_CSV`
- `HEAVY_FAV_THRESHOLD` (default: 1.40)
- `SAMPLE_SIZE` (optional)
- `SEED` (default: 42)

I. Next Steps

Potential extensions of the baseline pipeline include time-aware train-test splits; richer contextual features such as team strength, recent form, injuries, and rest days; strategy backtesting that incorporates transaction costs and liquidity constraints; and bookmaker-level modeling with robust aggregation methods.

VII. Limitations

This study relies on aggregated bookmaker odds, which provide a stable representation of market consensus but may obscure informative differences across individual bookmakers. The analysis is restricted to the Big Five European leagues, limiting the generalizability of results to smaller or less liquid markets. In addition, match outcomes are modeled using implied probabilities derived solely from odds, without incorporating football-specific contextual variables. As a result, the predictive models should be interpreted as baselines rather than fully optimized forecasting systems.

VIII. Future Work

Future research may extend this framework by introducing time-aware train–test splits and richer contextual features such as team strength, recent form, and injuries. Analyzing bookmaker-level odds instead of aggregated prices may provide further insight into market dynamics. Finally, strategy-based backtesting with realistic constraints could help assess whether observed deviations are economically exploitable.

References

[1] E. B. Cansever, "Whistles or whispers," GitHub repository, 2026, accessed: 2026-01-09. [Online]. Available: https://github.com/erencansever/Whistles-or-Whispers