NATIONAL ECONOMICS UNIVERSITY
SCHOOL OF TECHNOLOGY

Data Preparation and Visualization

Final project

Decoding the Cinderella Effect: A Data Preparation and Visualization Approach to Explaining Football Upsets or Cinderella effect

---

Advisor:	Nguyen Tuan Long	
Students:	Nguyen Viet Hoang	11230539
	Do Anh Ly	11230563
	Hoang Thi Thanh Nhan	11230578
	Tran Dinh Tuan Phong	11230581
	Truong Hoang Tung	11230601

HA NOI CITY, DECEMBER 2025

Table of contents

1 Executive summary

This project uses the power of data preprocessing and visualization to analyze the "Cinderella" phenomenon in football—teams that exceed expectations despite low initial ratings. By cleaning, standardizing, and visualizing complex football data, the project makes hidden performance patterns easier to detect. This is essential for understanding what drives unexpected success and for improving analysis, scouting, and prediction in modern football.

2 Introduction and Data Context

2.1 The Cinderella Archetype

2.1.1 The Literary Metaphor

Universally recognized, the tale of *Cinderella* by Charles Perrault serves as the quintessential archetype of unexpected triumph. It depicts the journey of a poor girl, relegated to the shadows and dismissed by the established order. However, through a profound metamorphosis, her humble rags are reconstructed into royal attire, enabling her to transcend her station. In a single decisive night, she steps out of obscurity to defy the hierarchy and capture the spotlight. This narrative endures not merely as a fairy tale, but as the ultimate symbol of the "underdog"—the improbable rise from the bottom to the very top.

2.1.2 Cinderella in Sports

In the context of sports, the term "Cinderella" denotes a low-rated underdog that secures a victory against a superior opponent or achieves success far beyond statistical expectations. A quintessential example is Leicester City's 2015-2016 Premier League title campaign. Despite possessing limited resources and facing 5000-to-1 betting odds at the season's start, the team outperformed financially dominant rivals to win the championship. This case demonstrates that competitive outcomes can deviate significantly from static, resource-based predictions.

3 Data storytelling

3.1 Set Up

3.1.1 Project context

Football fans, researchers, coaches, and analysts are always interested in match outcomes and team strategies. Normally, teams with a higher-rated (the Favorite) are expected to defeat lower-rated teams (the Underdog), which is unsurprising. However, our project focuses on the rare events when underdog teams unexpectedly defeat strong teams, despite having very low pre-match winning probabilities. These events are referred to as upset matches or the Cinderella Effect.

3.1.2 Dataset Information

Our dataset was sourced from the Github platform, covering matches from 2000 to 2025. Because smaller leagues often lack reliable data, we focused on the five largest European leagues: Ligue 1 (F1), Bundesliga (D1), Premier League (E0), La Liga (SP1), and Serie A (I1). From the original 48 features, we selected 29 relevant features for this project. The final dataset thus contains 43,708 rows and 29 columns, and the data is evenly distributed across leagues to avoid bias toward any single league.

Table 3.1: Essential Features (29 Variables)

Features	Data Types	Description
Division	Enum	League that the match was played in - country code + division number (I1 for Italian First Division). For countries where we only have one league, we use 3-letter country code (ARG for Argentina).
MatchDate	Datetime	Match date in the classic YYYY-MM-DD format.
MatchTime	Datetime	Match time in the HH:MM:SS format. CET-1 time zone.
HomeTeam	String	Home team's club name in English, abbreviated if needed.
AwayTeam	String	Home team's club name in English, abbreviated if needed.
HomeElo	Float	Measures team strength dynamically; key for predicting favorites and underdogs. Home team's most recent Elo rating.

Table 3.1: Essential Features (Continued)

Features	Data Types	Description
AwayElo	Float	Measures team strength dynamically; key for predicting favorites and underdogs. Away team's most recent Elo rating.
Form3Home	Int	Number of points gathered by home team in the last 3 matches (Win = 3 points, Draw = 1 point, Loss = 0 points, so this value is between 0 and 9).
Form5Home	Int	Number of points gathered by home team in the last 5 matches (Win = 3 points, Draw = 1 point, Loss = 0 points, so this value is between 0 and 15).
Form3Away	Int	Number of points gathered by away team in the last 3 matches (Win = 3 points, Draw = 1 point, Loss = 0 points, so this value is between 0 and 9).
Form5Away	Int	Number of points gathered by away team in the last 5 matches (Win = 3 points, Draw = 1 point, Loss = 0 points, so this value is between 0 and 15).
FTHome	Int	Full-time goals scored by home team.
FTAway	Int	Full-time goals scored by away team.
FTResult	Enum	Full-time result (H for Home win, D for Draw and A for Away win).
HTHome	Int	Half-time goals scored by home team.
HTAway	Int	Half-time goals scored by away team.
HTResult	Enum	Half-time result (H for Home win, D for Draw and A for Away win).
HomeShots	Int	Total shots (goal, saved, blocked, off-target) by home team.
AwayShots	Int	Total shots (goal, saved, blocked, off-target) by away team.
HomeTarget	Int	Total shots on target (goal, saved) by home team.
AwayTarget	Int	Total shots on target (goal, saved) by away team.

Table 3.1: Essential Features (Continued)

Features	Data Types	Description
HomeFouls	Int	Total fouls by home team. (Aggressiveness metric; may influence game control, yellow/red card prediction.)
AwayFouls	Int	Total fouls by away team. (Aggressiveness metric; may influence game control, yellow/red card prediction.)
HomeCorners	Int	Total corners taken by home team. (Often correlated with attacking pressure; useful in offensive analysis.)
AwayCorners	Int	Total corners taken by away team. (Often correlated with attacking pressure; useful in offensive analysis.)
HomeYellow	Int	Yellow cards for home players (excluding staff)
AwayYellow	Int	Total yellow cards awarded to away team players (excl. staff).
HomeRed	Int	Total red cards awarded to home team players (excl. staff).
AwayRed	Int	Total red cards awarded to away team players (excl. staff).

3.1.3 Project targets

Initially, we see some problems in raw features include noise, inconsistent recording across leagues, and variables not directly informative for predicting upsets. These issues make it difficult for a model to detect hidden patterns in raw data and emphasize the need for careful preprocessing. To improve model performance, we apply data preprocessing. Our primary objective is not to find the model for the best performance because predict rare events is really difficult, instead of that we aim to build an efficient predictive model that can identify potential Cinderella outcomes before the matches occur.

In predictive modeling, our central hypothesis is whether these anomalies represent mere 'random noise' or indicate a 'latent pattern' governed by dynamic factors not captured in the dataset. To quantify team strength and establish a baseline expectation for match outcomes, we use the Elo Rating System, which calculates team ratings based on historical results. This allows us to compute the A Priori Win Probability (P) of an underdog winning against a favorite using the formula:

$$P = \frac{1}{1 + 10^{\frac{\text{Elo}_B - \text{Elo}_A}{400}}} \tag{3.1}$$

So, our target is defined as `is_cinderella` when a match satisfies 2 conditions:

$$\begin{cases} P \leq 0.3 \\ \text{FTResult} = \text{H or A} \end{cases}$$

Our definition specifies that a match is designated for analysis only if the underdog team's A Priori Win Probability is less than 30 percent. After that, we make a comparison the effects of raw vs processed data on model performance. Finally, we use analysis interpretability methods such as feature importance and SHAP values to understand which factors contribute most to upsets. After this stage, we found out that the number of matches are rare upsets or appeared with a Cinderella effect.

3.2 Conflict

3.2.1 Initial exploratory analysis with raw dataset

Our raw data have 43708 rows and 30 features (29 features originally and 1 target column)

	Division	MatchDate	MatchTime	HomeTeam	AwayTeam	HomeElo	AwayElo	Form3Home	Form5Home	Form3Away	Form5Away	FTHome	FTAway	FTResult	HTHome	HTAway	HTResult
0	F1	2000-07-28	NaN	Marseille	Troyes	1686.34	1586.57	0.0	0.0	0.0	0.0	3.0	1.0	H	2.0	1.0	H
1	F1	2000-07-28	NaN	Paris SG	Strasbourg	1714.89	1642.51	0.0	0.0	0.0	0.0	3.0	1.0	H	1.0	1.0	D
3	F1	2000-07-29	NaN	Auxerre	Sedan	1635.58	1624.22	0.0	0.0	0.0	0.0	0.0	1.0	A	0.0	1.0	A
4	F1	2000-07-29	NaN	Bordeaux	Metz	1734.34	1673.11	0.0	0.0	0.0	0.0	1.0	1.0	D	1.0	0.0	H
5	F1	2000-07-29	NaN	Guingamp	St Etienne	1578.51	1620.74	0.0	0.0	0.0	0.0	2.0	2.0	D	2.0	1.0	H

HomeShots	AwayShots	HomeTarget	AwayTarget	HomeFouls	AwayFouls	HomeCorners	AwayCorners	HomeYellow	AwayYellow	HomeRed	AwayRed
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 3.1: Features of raw data

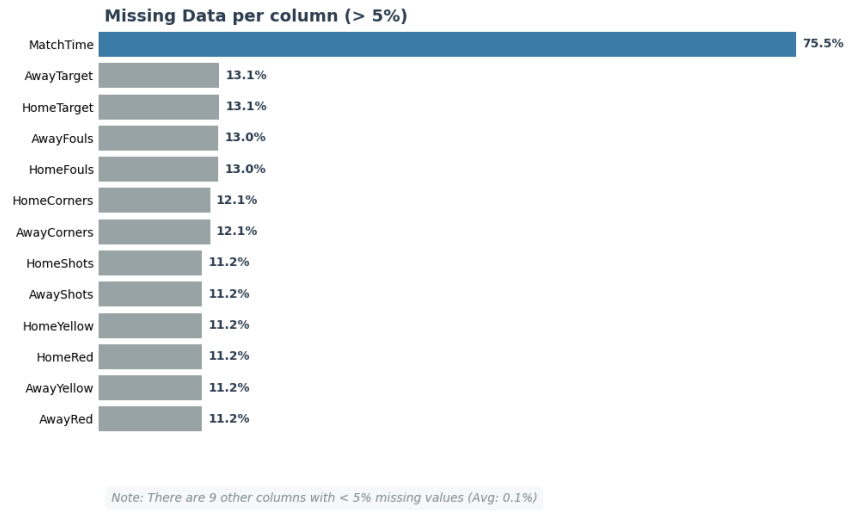


Figure 3.2: Percentage of Missing Values per Columns

Raw data lacks lots of data, 75% for 'MatchTime', any around 10% for 12 other columns and just 5% for remaining cols in the chart.

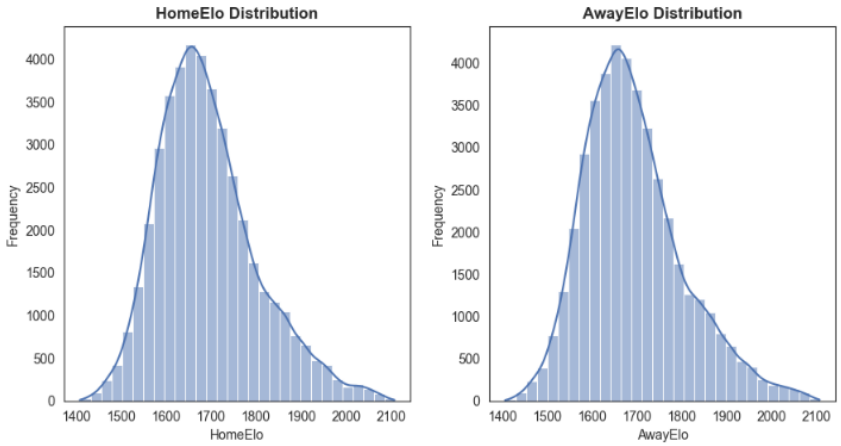


Figure 3.3: Elo Distribution

Look at these charts, all of distribution show that there are real values in those outliers not noise because all values are still valid and in possible range. These outliers represent legitimate extreme states in football dynamics rather than data collection errors In data science terms, these are 'phenomenological outliers'—rare but valid events that carry the strongest signals for our target variable. Treating them as noise to be removed would strip the dataset of the very high-disparity scenarios where 'Cinderella' upsets are most likely to occur After identifying the target, we find out that the number of matches that has the Cinderella effect is:

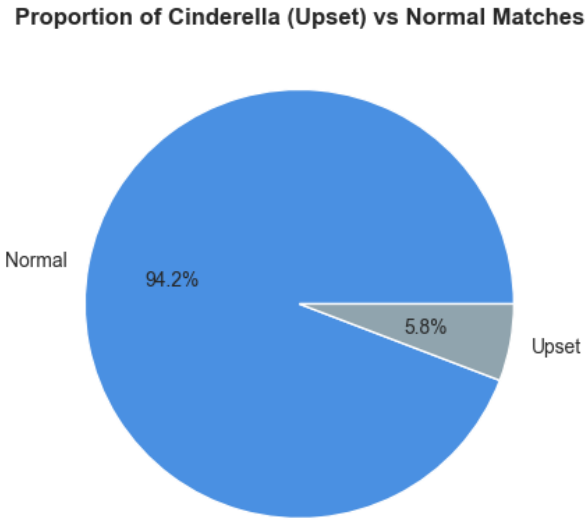


Figure 3.4: Proportion of Cinderella vs Normal Matches

We found that only 5.8% of matches qualify as Cinderella. This heavy class imbalance makes prediction extremely challenging, as the signal we want to detect occupies just a tiny fraction of the data. Hence, it's very difficult to predict or detect the effect.

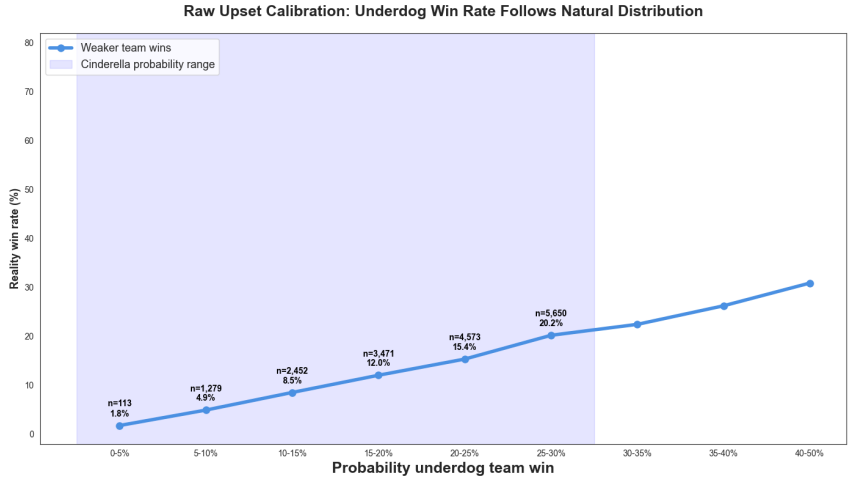


Figure 3.5: Cinderella Rate by Underdog Location

Raw upset calibration behaves as expected: underdog win rate increases smoothly with predicted probability, showing a natural statistical pattern. In the low-probability region (<30%), cinderella rate equals weak team win because Cinderella is simply a subset of upsets defined by a fixed threshold. Thus, Cinderella does not exhibit any independent pattern in raw data — it is just part of the noisy tail. Conclusion: detecting Cinderella requires preprocessing and feature engineering to extract meaningful signal.

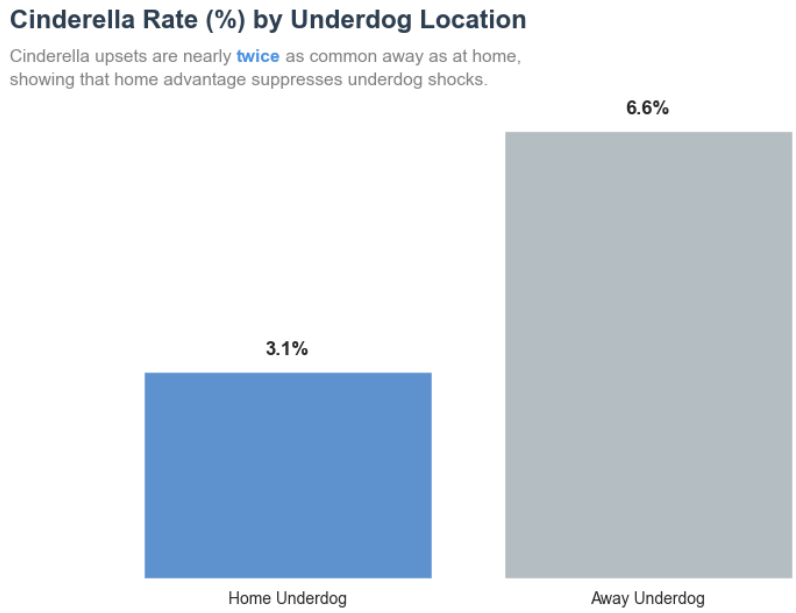


Figure 3.6: Underdog win rate (even cinderella) calibration

Further analysis reveals that Cinderella events are significantly rarer when the underdog plays at home (3.1%) compared to away (6.6%). This highlights the role of home advantage — even for underdogs, playing at home makes it harder to pull off an upset.

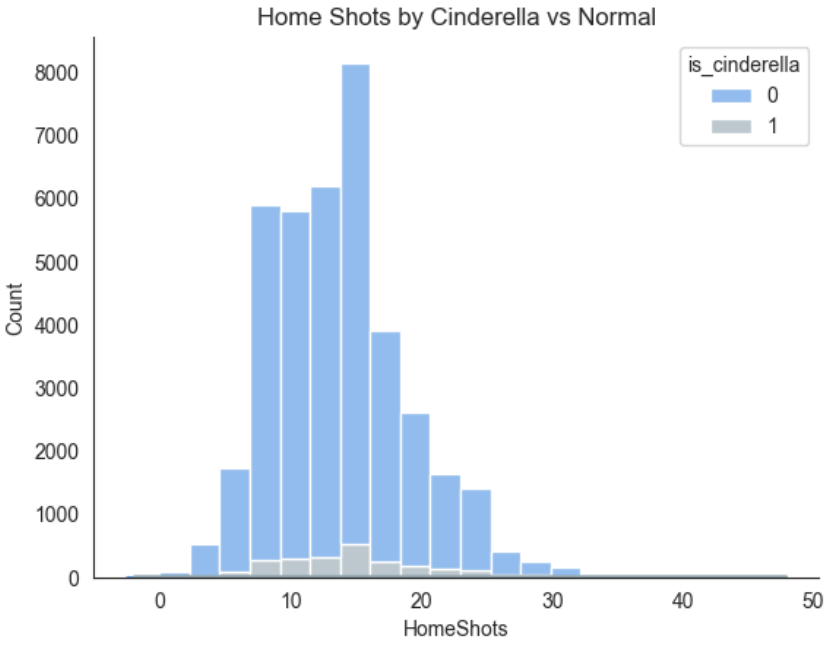


Figure 3.7: Home Shots by Cinderella vs Normal

Even when a massive Cinderella upset occurs, the underdog takes virtually the same number

of shots as in a normal match. Both distributions peak at 12–16 shots and are nearly identical. The right tail (>25 shots) is dominated by non-Cinderella matches. Average Home Shots in Cinderella matches is only 0.5–1 shot higher than usual — statistically indistinguishable. Most people intuitively assume that huge upsets are accompanied by exceptional attacking stats from the weaker team.

3.2.2 Uncovering Hidden Signals — Where the Real Patterns Lie

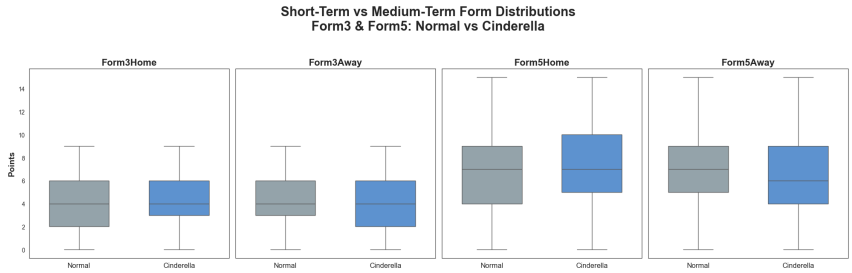


Figure 3.8: Form Distributions (Normal vs Cinderella)

Contrary to initial visual impressions from the boxplots — which suggest subtle differences in distribution between Cinderella and Normal groups — the Mann-Whitney U tests reveal a statistically significant difference for all four form features (Form3Home, Form3Away, Form5Home, Form5Away), with p-values consistently at 0.0000. This apparent contradiction highlights a crucial insight: while medians are identical (e.g., Form5Away: 7.0 vs 6.0; others: 4.0 vs 4.0 or 7.0 vs 7.0), the test is not comparing medians alone — it evaluates the entire distribution, including spread, skewness, and outlier behavior. The fact that p-values are near zero despite identical medians suggests that the 'signal' lies not in central tendency, but in how values are distributed around it — for example, Cinderella events may exhibit narrower interquartile ranges, fewer extreme outliers, or different tail behaviors. This reinforces our core argument: raw data can be deceptive. This is why we cannot rely solely on correlation or summary statistics — and why feature engineering must go beyond simple aggregations to capture these hidden distributional patterns, such as variance ratios, percentile gaps, or interaction terms with Elo or location.

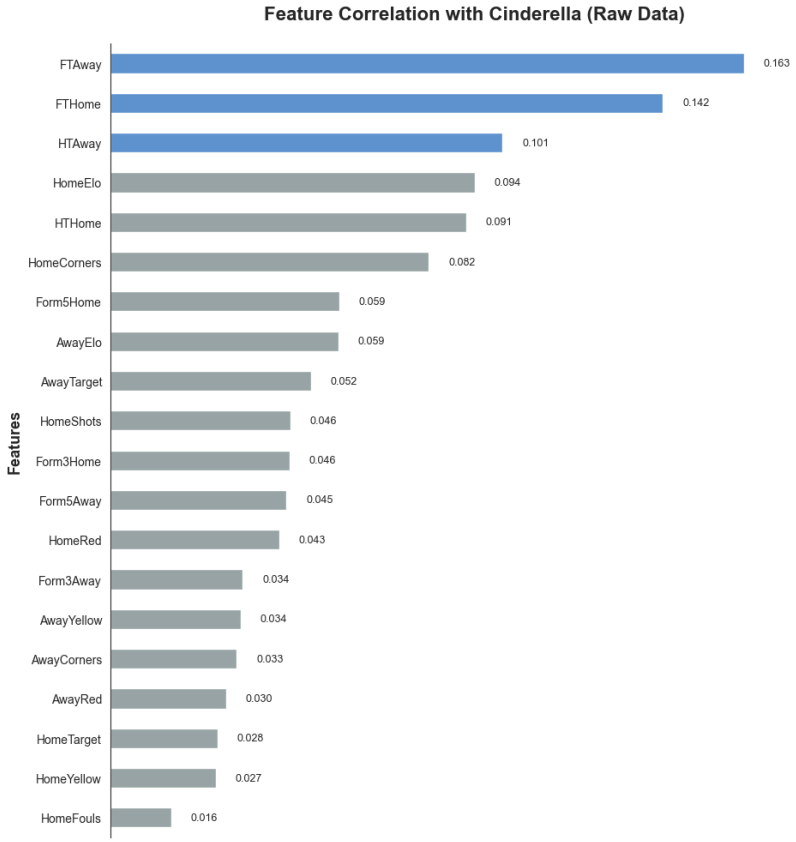


Figure 3.9: Feature Correlation with Cinderella (Raw Data)

The correlation heatmap reveals a critical truth about our dataset: while raw features like final scores (FTAway, FTHome) show the highest absolute correlations with Cinderella events (0.163 and 0.142 respectively), even these are remarkably weak. All other variables — including team strength (Elo), in-game metrics (shots, corners, cards), and short/medium-term form (Form3, Form5) — exhibit correlations below 0.06, with many hovering near zero (e.g., HomeFouls: 0.016). This is not an accident — it is the signature of a rare, noisy, and inherently unpredictable phenomenon. The lack of strong linear relationships confirms that Cinderella upsets cannot be predicted using raw, unprocessed features. Even the most 'obvious' signals — such as goals scored or Elo ratings — are too blunt to capture the subtle, non-linear dynamics that define true underdog victories. This explains why earlier visualizations (boxplots) and statistical tests (Mann-Whitney U) were necessary: they revealed distributional differences invisible to correlation analysis. In essence, correlation tells us what's not predictive — and that's precisely why we must move beyond it. The path forward lies in feature engineering: transforming raw inputs into derived metrics that encode context, interaction, and hidden patterns — turning noise into signal, and illusion into insight.

3.2.3 Baseline model and problems

Before training any predictive model, we first addressed the issue of missing values in the raw dataset. Since the proportion of missingness was substantial and affected multiple key variables, we opted to remove all rows containing missing entries. This resulted in a cleaned dataset of 32,964 observations.

To ensure a fair evaluation and prevent any form of data leakage, we restricted the feature set to strictly pre-match variables. These include:

- `HomeElo`
- `AwayElo`
- `Form3Home`
- `Form3Away`
- `Form5Home`
- `Form5Away`

These features represent team strength and short-term form prior to kickoff, ensuring that no post-match information influences the training process.

For modeling, we selected a single tree-based algorithm — **LightGBM** — and applied it consistently to both raw-feature and engineered-feature scenarios. This choice allows us to isolate the impact of preprocessing and feature engineering on predictive performance while keeping the model architecture fixed.

```
RESULTS
  Best Threshold : 0.760
  Test F1-Score  : 0.2500
  ROC-AUC        : 0.8009
  PR-AUC         : 0.1648

  Classification Report:
              precision    recall  f1-score   support

           0     0.9661    0.7800    0.8631      6136
           1     0.1578    0.6010    0.2500       421

    accuracy                         0.7685      6557
   macro avg     0.5620    0.6905    0.5566      6557
weighted avg     0.9142    0.7685    0.8238      6557
```

Figure 3.10: Metrics Report of LightGBM Model

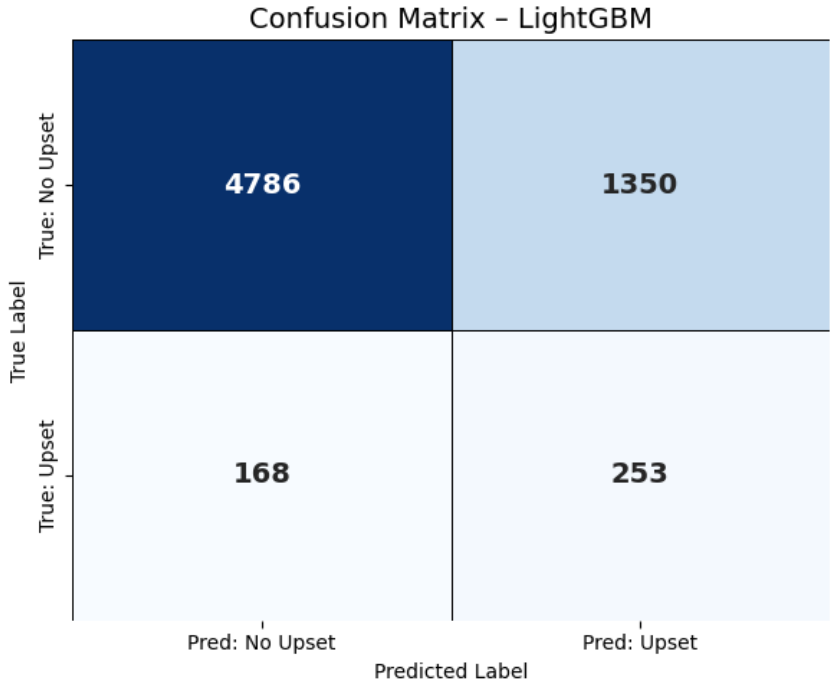


Figure 3.11: Confusion Matrix of LightGBM Model with Raw Data

Cinderella matches are rare ($\sim 5.8\%$), creating a heavily imbalanced dataset. The model reflects this imbalance: it achieves high precision and recall for the majority class (normal matches) but struggles with low precision for the Cinderella class.

- **F1-score (Cinderella class):** 0.25
- **Recall:** 0.60 — the model captures more than half of the rare Cinderella events, indicating that even raw features contain a weak but detectable signal.
- **ROC-AUC:** 0.80 — the classifier demonstrates a solid ability to rank positive instances above negative ones.
- **PR-AUC:** 0.16 — low, reflecting the inherent difficulty of detecting rare events under severe class imbalance.

Surprisingly, the raw-data model is still able to detect a portion of Cinderella events despite minimal feature processing. This indicates that even without engineered variables, the dataset contains latent patterns that differentiate true upsets from normal matches, although these patterns remain weak and highly noisy.

Analysis of the confusion matrix shows that the model successfully captures a considerable share of rare upsets, reinforcing the existence of underlying signals within the raw features. Most misclassifications arise from predicting normal matches as Cinderella events (False Positives),

suggesting that the model adopts a conservative decision boundary but continues to struggle with the severe class imbalance.

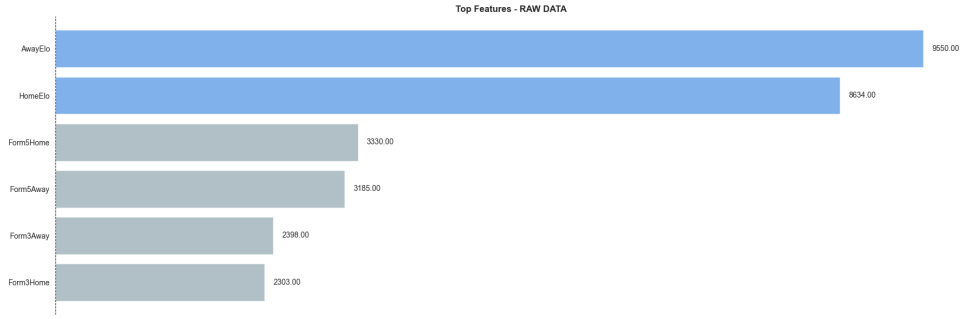


Figure 3.12: Feature Importance of Raw Data

The feature importance plot for the baseline model shows a clear hierarchy in predictive power. Elo-based variables dominate the ranking, with `AwayElo` achieving an importance score of 9550 and `HomeElo` following at 8634. In contrast, recent form metrics (`Form3Home`, `Form3Away`, `Form5Home`, `Form5Away`) occupy a much lower range, between roughly 2300 and 3300.

This pronounced gap indicates that, in the raw-feature setting, the model relies predominantly on long-term team strength captured by the Elo Ratings, while short-term performance fluctuations contribute substantially less to the prediction process.

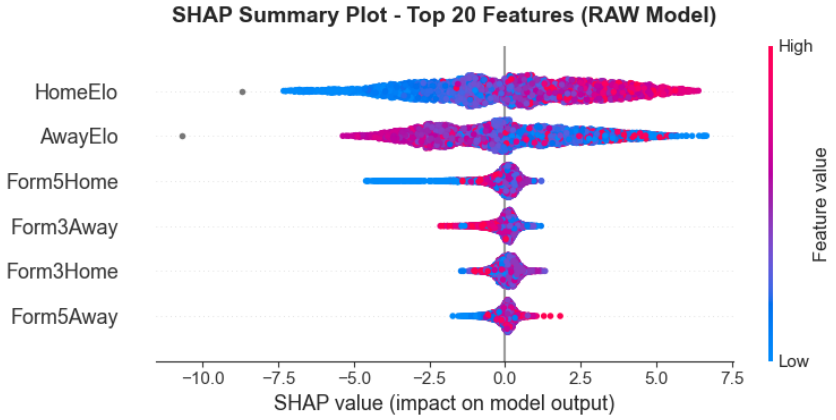


Figure 3.13: SHAP Summary Plot of Raw Data

The SHAP summary plot provides further insight into the directionality and magnitude of these features:

- **Elo Ratings:** High values of `HomeElo` (red points) correspond to positive SHAP values, pushing predictions toward the target class (higher likelihood of a home win). In contrast, high `AwayElo` values produce negative SHAP contributions, reflecting the intuitive effect that a stronger away team lowers the home team's predicted win probability.

- **Form Features:** Variables such as `Form5Home`, `Form3Away`, and others cluster tightly around zero. Although their effects show consistent directional tendencies (e.g., higher `Form5Home` slightly increases prediction scores), their overall influence is small relative to the Elo ratings.

Finally, even without extensive preprocessing, the raw dataset still contains enough signal for the LightGBM model to identify Cinderella matches at a level better than random guessing. This supports the hypothesis that hidden patterns do exist within the unprocessed features and that these signals can be better leveraged through feature engineering, preprocessing, and well-designed derived variables to uncover clearer structure and improve predictive performance.

3.3 Resolution

3.3.1 Feature Engineering

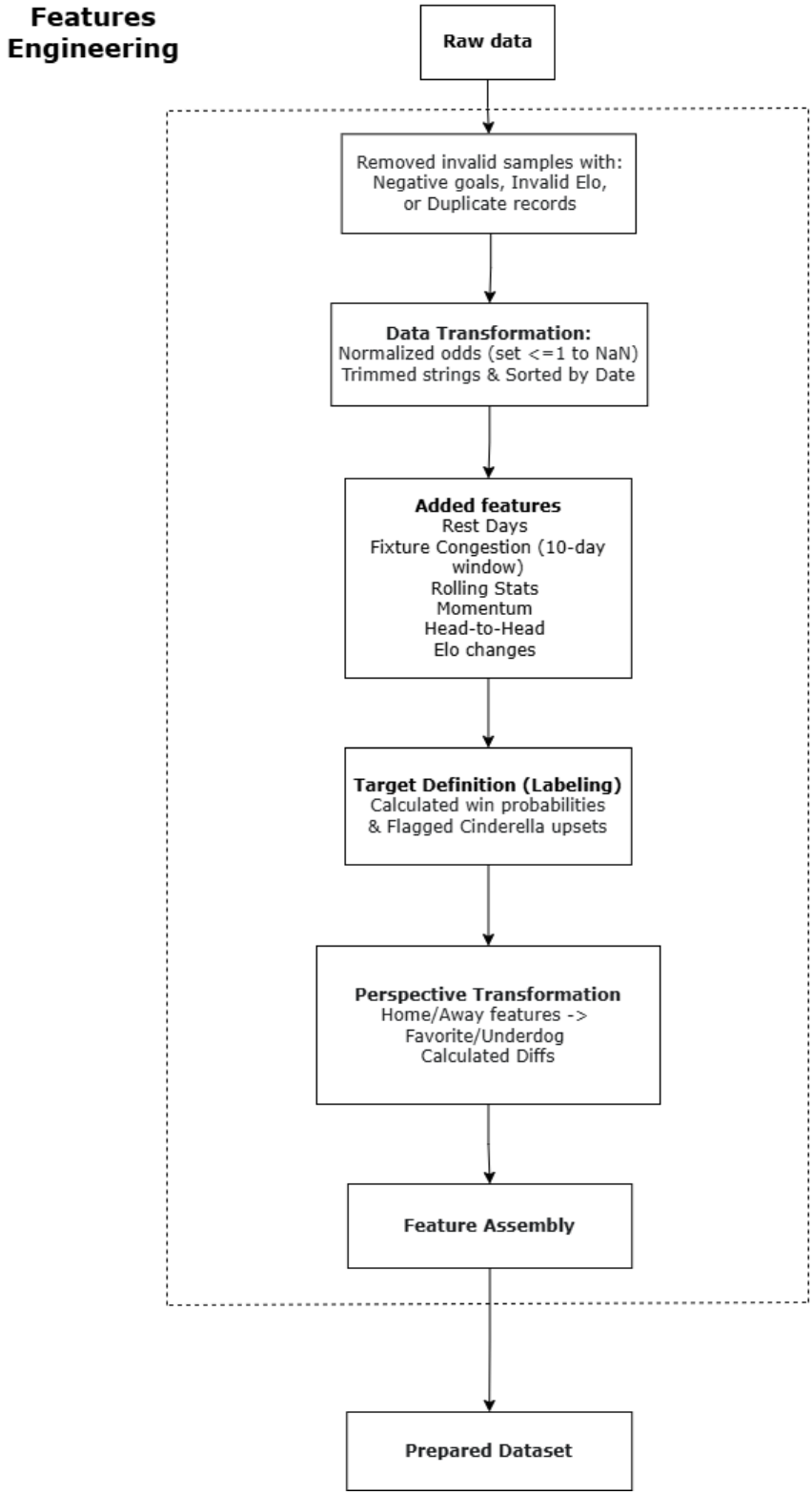


Figure 3.14: Feature engineering flowchart

a. Data Cleaning Pipeline

Objective The Data Cleaning Pipeline establishes the **structural integrity** required for robust predictive modeling. This procedure subjects the raw input to a series of validation filters, transforming unstructured data into a consistent analytical schema. The removal of erroneous artifacts prevents the downstream propagation of logical fallacies into feature engineering and algorithmic training.

Methodological Specification The process initiates with **Temporal Standardization**, converting `MatchDate` string literals to **datetime objects** to ensure chronological precision for time-series derivation. Simultaneously, **Redundancy Elimination** is enforced via `drop_duplicates()` to remove identical feature vectors, ensuring the **statistical independence** of observations and preventing the model from memorizing repetitive patterns.

Following these initial steps, **Logical Verification** enforces fundamental domain constraints by excising observations that violate physical and statistical plausibility. Negative values for goal counts (`FTHome`, `FTAway`) represent logical impossibilities, while Elo ratings of zero indicate corrupted external data that render the observation mathematically undefined for probability calculation.

⇒ This guarantees **data plausibility** and preserves the validity of the vector space.

Listing 3.1: Logical Verification Code

```
# 1.2 Remove invalid data
df_prepared = df_prepared[
    (df_prepared['FTHome'] >= 0) &
    (df_prepared['FTAway'] >= 0) &
    (df_prepared['HomeElo'] > 0) &
    (df_prepared['AwayElo'] > 0)
]
```

⇒ This enhances **distributional stability** and reduces the model's sensitivity to **stochastic noise** during convergence.

b. Feature Engineering Pipeline

Table 3.2: Comprehensive Feature Engineering Specifications

Variable Name	Formula (Logic)	Operational Meaning
Rest Days (`RestDaysHome/Away`)	$\Delta t = \text{Date}_t - \text{Date}_{t-1}$ *(Default = 7 days if n/a)*	Measures acute **biological recovery**. Intervals ≤ 3 days indicate high physical stress.
Fixture Congestion (`Congestion_10d`)	$\sum \text{Matches} \in [\text{Date}_t - 10, \text{Date}_t)$ *(Strictly < Current Date)*	Quantifies **chronic fatigue** due to schedule density, independent of the single previous match.
Rolling Form (`Points_Rolling`)	$\sum_{i=1}^{k} \text{Points}_{t-i}$	Captures **psychological momentum** and short-term consistency.
Rolling Goals (`GF/GA_Rolling`)	$\sum \text{GF}_{t-i}$ (Offense) $\sum \text{GA}_{t-i}$ (Defense)	Assesses actual **offensive/defensive output**, highlighting structural imbalances.
Rolling Tactics (`Shots, Target, Corners`)	$\text{Avg}(\text{Metric\_For}_{t-i})$ vs $\text{Avg}(\text{Metric\_Conceded}_{t-i})$	Measures underlying **game dominance** (xG proxy) and defensive suppression capabilities.
Rolling Discipline (`Fouls, Cards`)	$\sum \text{Cards}_{t-i}$ *(for $i = 1..k$)*	Proxies for aggression levels and potential suspension risks.
Momentum	$(\text{Form}_3/3) - (\text{Form}_5/5)$	Measures the **acceleration of performance**. Positive values indicate an improving trend.
Head-to-Head (H2H) (`H2H_Wins, Points`)	$\sum \text{Outcomes}(\text{Team A vs B})_{\text{past}}$	Encodes specific **matchup psychology** and historical rivalry dominance.
Elo Change (`EloChange1/2`)	$\text{Elo}_t - \text{Elo}_{(t-30 \text{ days})}$	Quantifies **structural evolution** (improvement or decline) over medium-term horizons.

Note: *All rolling calculations (where $k = 3, 5$) utilize a shifted index ($t - 1$) to ensure zero data leakage. H2H features strictly exclude the current match outcome.*

c. Target Definition & Comparative Feature Construction

Objective This phase fundamentally restructures the dataset from a geographical perspective (Home/Away) to a **competitive perspective** (Favorite/Underdog). By explicitly defining the target anomaly and realigning features relative to team strength, the model is forced to learn the specific dynamics of "upset" events rather than generic match outcomes.

Methodological Specification **1. Probabilistic Target Derivation (The Cinderella Label)**

Declarative Claim: The binary target variable `is_cinderella` is constructed via a strict probabilistic thresholding mechanism based on the Elo Rating System.

Justification: A generic "Underdog Win" is insufficient for anomaly detection, as it includes matches with near-equal odds (e.g., 49% vs 51%). To isolate true "Cinderella" events, we must identify victories that were statistically improbable.

Specification:

- **Step A:** Calculate Win Probability (P_{win}) using the logistic function of the Elo difference, adjusted for Home Advantage.
- **Step B:** Classify the Underdog as the entity with $P_{win} < 0.5$.
- **Step C:** Assign the target label $Y = 1$ if and only if the Underdog wins the match **AND** the Underdog's pre-match probability was below the defined `PROB_THRESHOLD`.

$\Rightarrow$ Converts the problem into a **constrained binary classification task** focused on high-variance anomalies.

Listing 3.2: Target Creation Logic

```
# Target Creation Logic
df_prepared['Elo_diff'] = (
    df_prepared['HomeElo'] + HOME_ADVANTAGE - df_prepared['
       AwayElo']
)
df_prepared['Prob_HomeWin'] = 1 / (1 + 10 ** (-df_prepared['
   Elo_diff'] / 400))

# ... (Favorite/Underdog identification) ...
df_prepared['is_cinderella'] = (underdog_wins & is_low_prob).
   astype(int)
```

2. Feature Space Re-orientation (Vectorization)

Declarative Claim: All predictive features are transformed from absolute Home/Away attributes into relative **Favorite/Underdog vectors**.

Justification: Raw features such as `HomeShots` are context-dependent; a high shot count for a Favorite is expected, whereas a high shot count for an Underdog signals a potential upset. Re-orienting the feature space ensures the model evaluates performance relative to expectation.

Specification:

- **Conditional Mapping:** Utilizing vectorized `np.where` logic, attributes (Elo, Rest Days, Momentum, Form, H2H) are remapped to `_fav` and `_underdog` suffixes.
- **Differential Features:** New features are generated to quantify the competitive gap (e.g., `Elo_diff`, `Momentum_diff`, `Form3_diff`).

⇒ Enhances **class separability** by explicitly encoding the asymmetry between competitors.

Listing 3.3: Feature Vectorization Code

```
# Example: Creating relative features
df_prepared['Momentum_underdog'] = np.where(
    is_home_underdog,
    df_prepared['MomentumHome'],
    np.where(is_away_underdog, df_prepared['MomentumAway'],
        np.nan)
)
df_prepared['Momentum_diff'] = (
    df_prepared['Momentum_fav'] - df_prepared['
        Momentum_underdog']
)
```

3. Granular Technical Aggregation

Declarative Claim: The comparative transformation is extended to advanced rolling statistics (Shots, Corners, Fouls, Cards) and their conceded equivalents.

Justification: While macro-variables (Elo, Form) capture long-term trends, specific "Cinderella" signals often reside in micro-performance indicators (e.g., an Underdog conceding fewer shots on target than expected over the last 3 games).

Specification: The pipeline iterates through technical metrics to create a comprehensive set of **Performance Differentials** (`_diff`).

⇒ Provides the model with **high-resolution signals** regarding tactical efficiency and defensive solidity.

3.3.2 EDA on preprocessed data

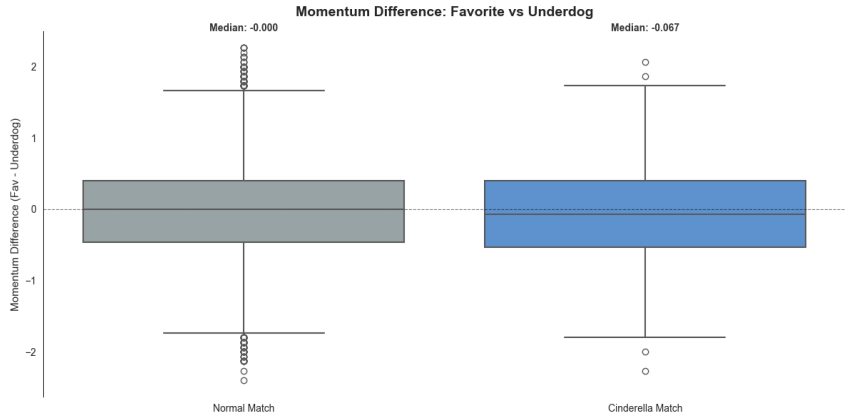


Figure 3.15: Momentum Difference: Favorite vs Underdog

The box plot provides compelling evidence that Cinderella events are not random anomalies but are driven by superior underdog form. While *Normal Matches* exhibit perfect parity in momentum between teams (Median: 0.000), *Cinderella Matches* display a distinct negative deviation (Median: -0.067).

This shift reveals a critical latent pattern: upsets are significantly more likely when the underdog enters the match with greater momentum than the favorite. This validates Momentum Difference as a high-value discriminator, successfully transforming a raw concept into a quantifiable predictive signal.

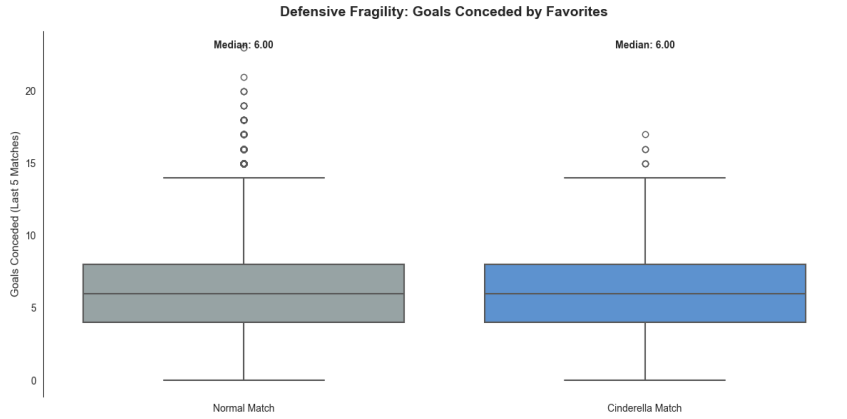


Figure 3.16: Defensive Fragility: Goals Conceded by Favorite

Contrary to the expectation that "leaky defenses" directly invite upsets, the data reveals identical medians (6.00) for favorites in both Normal and Cinderella matches. This result highlights a critical distinction: defensive fragility alone is not a sufficient trigger for an upset. The lack of univariate separation reinforces the necessity of our machine learning approach, suggesting that a favorite's defensive weakness is contextual—only exploited when combined with specific high-risk factors, such as the superior underdog momentum identified earlier, rather than serving as an independent predictive signal.

3.3.3 Training model and result comparison

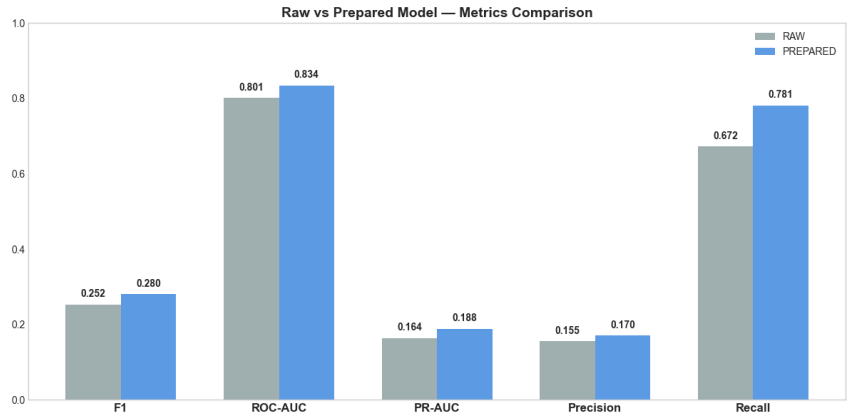


Figure 3.17: Raw vs Prepared Model - Metrics Comparison

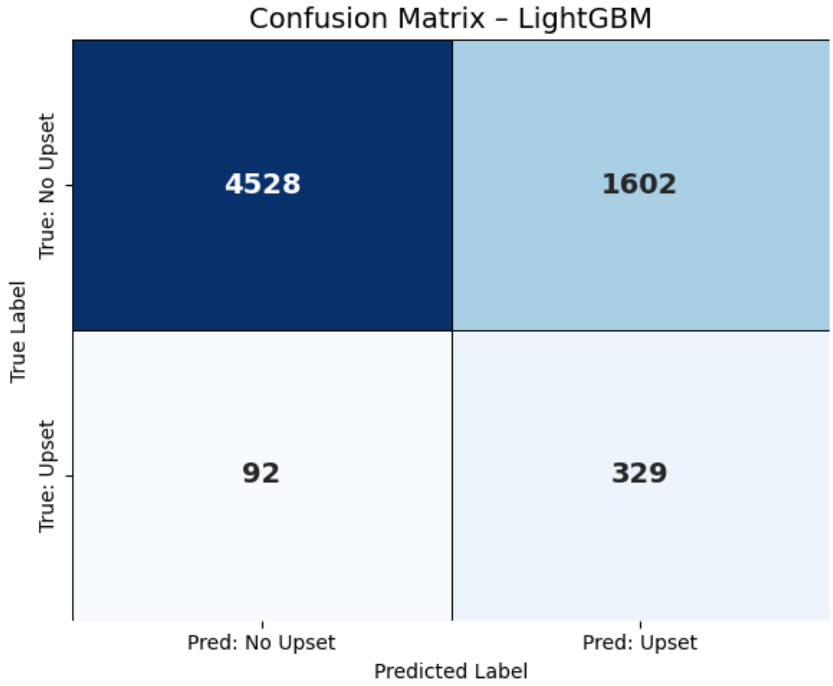


Figure 3.18: Confusion Matrix of LightGBM Model with Prepared Data

The transition from raw to prepared data yielded a universal improvement across all evaluation metrics, confirming our hypothesis that explicit feature engineering is necessary to "unlock" the latent signals associated with rare upset events.

The most significant achievement of the data preparation phase is the 16.3% increase in Recall (Sensitivity), rising from 0.672 to 0.781. This allows the model to capture nearly 80% of all *Cinderella* events. In applications where "missing the big game" constitutes the primary failure mode, this improvement serves as a key validation of our preprocessing strategy. Crucially, the gain in Recall was not due to indiscriminate prediction of "Upset" outcomes, but reflects genuine enhancement in model sensitivity.

Additionally, the Precision-Recall AUC (PR-AUC) improved by 14.5%, increasing from 0.164 to 0.188, demonstrating better identification of rare positive events under class imbalance.

F1-Score (+10.8%): The harmonic mean of precision and recall increased from 0.252 to 0.280, reflecting a stronger overall balance between detecting upsets and preserving predictive accuracy.

ROC-AUC (+4.1%): The increase to 0.834 demonstrates strong separability. The Prepared model has a superior ability to distinguish between a *Normal* match and a *Cinderella* match on a global scale.

By explicitly modeling team momentum and relative strength, we moved from a model that guessed based on reputation to a model that detects based on performance dynamics, achieving a substantial gain in its ability to foresee potential upsets.

3.3.4 Explanation of Cinderella effect

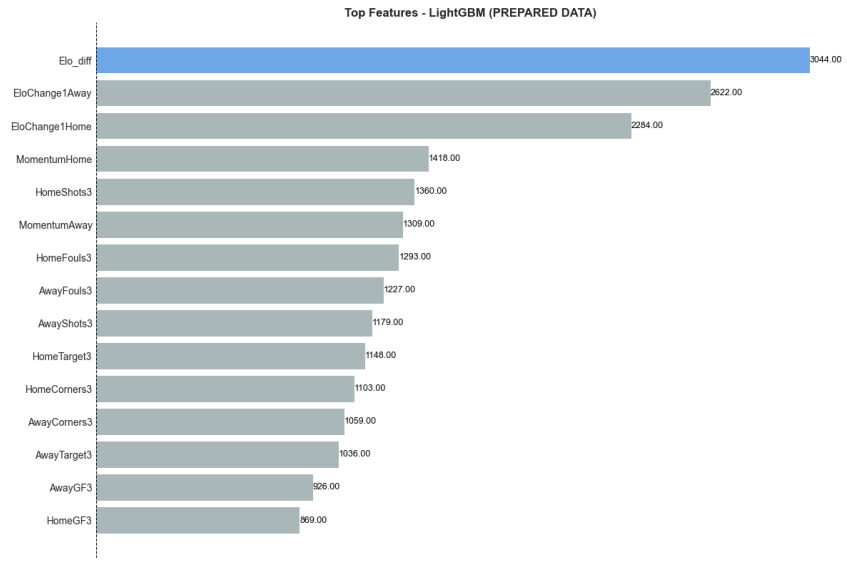


Figure 3.19: Feature Importance of Prepared Data

The LightGBM feature importance plot validates the efficacy of our feature engineering strategy, revealing a clear shift in how the model interprets match dynamics.

Dominance of Relative Strength (`Elo_diff`): The top predictor is no longer raw `Elo`, but `Elo_diff`. This indicates that the model now prioritizes the relative disparity between teams rather than their absolute strength, directly capturing the context necessary for identifying "David vs. Goliath" scenarios.

Capture of Recent Trends (`EloChange` & `Momentum`): Following closely are `EloChange` and `Momentum` features. Their high ranking confirms that the model is now sensitive to short-term form and trajectory, effectively distinguishing between a "stable" favorite and a "vulnerable" one based on recent performance fluctuations.

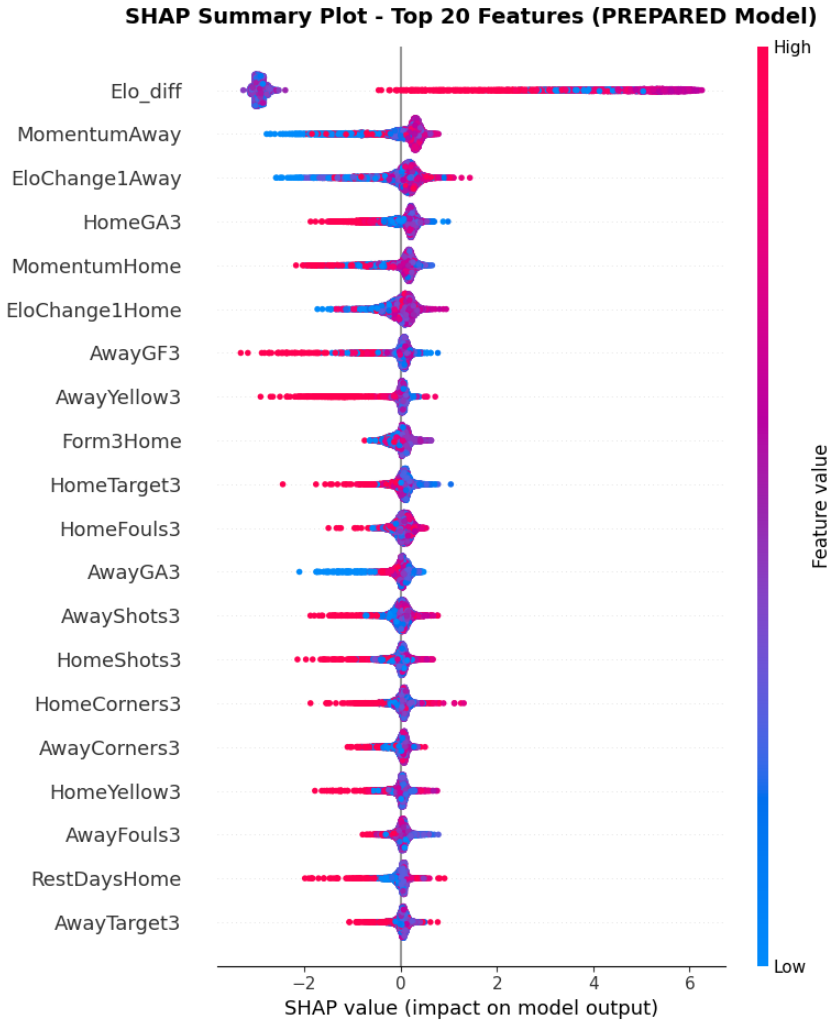


Figure 3.20: SHAP Summary Plot of Prepared Data

The SHAP analysis reveals that predicting upsets requires more than just the strength gap (`Elo_diff`). The model has successfully learned to identify a distinct "Cinderella" profile, driven by short-term performance dynamics.

High positive SHAP contributions from `MomentumAway` and `EloChange1_underdog` indicate that upsets are more likely when underdogs are actively improving and enter the match on an upward performance trajectory.

The prominence of 3-match rolling differences, particularly `GA3_diff` and `Corners3_diff`, confirms that favorites exhibiting recent defensive vulnerabilities or conceding tactical dominance (e.g., via set pieces) are more susceptible to upsets.

The model moves beyond static reputation, using engineered features to detect the specific interaction between an improving underdog and a stumbling favorite.

3.3.5 Conclusion

This study began with a central hypothesis: that football upsets are not merely stochastic anomalies (noise), but rather the result of latent patterns obscured by the high dimensionality of raw data.

Through targeted preprocessing and feature engineering, we successfully validated this hypothesis. By shifting the model's focus from static reputation (Raw `Elo`) to dynamic reality (`Momentum`, `Elo Change`, and tactical differences), we transformed the `Cinderella` phenomenon from an unpredictable outlier into a recognizable profile.

The quantitative improvements in Recall (+16.3%) and PR-AUC (+14.5%) confirm that explicitly modeling match context—particularly the interaction between a rising underdog and a vulnerable favorite—renders the previously "unpredictable" partially predictable.

4 Technical Analysis

4.1 Narrative Architecture and Strategic Framework

To effectively communicate our findings, this report combines two main storytelling strategies: **Chronological Ordering** and the **Three-Act Structure**. We primarily follow a Chronological path, which means presenting information in the order it occurred. As outlined in *Storytelling with Data*, this approach mirrors the actual steps of our analysis: identifying the problem, gathering data, analyzing it, and finding a solution. We chose this method to establish credibility with the audience. In a technical project, the reader needs to trust the process, not just the final result. By guiding the reader through our step-by-step journey, we demonstrate that our methods are reliable and our conclusions are based on solid evidence.

To make this timeline more engaging, we overlay the **Three-Act Structure** (Setup, Conflict, Resolution). We present the limitations of the raw data as a **Conflict**, creating tension by showing that standard methods fail to predict "Cinderella" events. This creates a "Call to Action" that naturally leads to the **Resolution**, where our new, engineered features appear as the solution to this conflict. By structuring the report this way, we transform a complex data problem into a satisfying story of discovery, keeping the audience interested from beginning to end.

4.1.1 ACT I: THE SETUP

To establish a clear context and piquing audience interest, we structure the project setup by addressing the five essential elements of strategic storytelling as outlined by Cliff Atkinson.

a. The Setting (Context & Scope)

Where and when does the story take place?

The analysis is situated in the domain of high-level European football, covering a 25-year timeline from **2000 to 2025**. Specifically, the dataset focuses on the "Big Five" leagues (Premier League, La Liga, Bundesliga, Serie A, Ligue 1) sourced from the Github platform. This scope serves as the analytical environment, comprising **43,708 observations** and **29 selected features**, ensuring that our investigation is grounded in a robust and representative sample of modern football history.

b. The Main Character (Who is driving the action?)

Who is the protagonist?

In this analytical narrative, the protagonist is the **Efficient Predictive Model**. Acting as a proxy for researchers and analysts, the model's goal is to decode the complex dynamics of match outcomes. Its mission is to look beyond the obvious "Favorite wins" scenarios and successfully identify value in high-risk situations.

c. The Imbalance (The Problem)

Why is this necessary? What is broken?

The current status quo is flawed. Initial inspection reveals that **raw data is inherently noisy** and inconsistent across leagues. Standard variables lack the informative power to detect subtle patterns. Consequently, traditional interpretations (like the standard Elo rating) often dismiss underdog victories as mere "Random Noise." This creates a state of **imbalance**: significant "Cinderella" events—where underdogs defeat favorites—are occurring, but our raw data tools are failing to predict them.

d. The Balance (The Desired Outcome)

What do we want to see happen?

We seek to restore balance by transforming unpredictable noise into interpretable signals. Our objective is **not** to achieve perfect accuracy (which is impossible for rare events), but to establish a **latent pattern**. We aim to reach a state where the model can proactively identify potential Cinderella outcomes *before* they occur, proving that these events are governed by dynamic factors rather than pure luck.

e. The Solution (The Methodology)

How will we bring about the changes?

The bridge between the current Imbalance and the desired Balance is **Feature Engineering** guided by the **Elo Rating System**.

- First, we use Elo to quantify the baseline expectation and mathematically define the "Underdog" (via A Priori Win Probability).
- Second, we implement rigorous preprocessing to clean the noise.
- Finally, we engineer dynamic variables to capture hidden drivers of performance.

4.1.2 ACT 2: THE CONFLICT

In Act II, we proceed with the standard Exploratory Data Analysis (EDA) on the raw data set. By continuing to 'Develop the Situation,' we drive the narrative toward a dramatic climax at the conclusion of the Conflict phase, explicitly exposing the fundamental structural issues and the inherent failures of the raw data approach.

a. Phase 1: Preliminary Analysis – Establishing the "Villain"

In the first phase, we intentionally characterize the Raw Data as the "Villain" of the story. By highlighting severe flaws—such as **75% missing values** in `MatchTime` and chaotic outliers—we demonstrate that the raw environment is unstable. This challenge is compounded by **class imbalance**, as "Cinderella" events occur in only **5.8%** of matches.

Narratively, this creates a "Crisis." It proves to the audience that standard predictive models will inevitably fail because the target signal is buried under too much noise. This establishes that a simple approach is insufficient.

b. Phase 2: Latent Pattern Discovery – The "Plot Twist"

Just as the analysis seems to reach a dead end, we introduce a narrative "Plot Twist." While visual charts suggest that winners and losers look identical, **statistical tests** reveal that hidden (latent) patterns do exist. However, the **Baseline Model** still fails to predict accurate results because it cannot interpret these subtle signals within the raw noise. This contrast serves a strategic purpose: it proves that the signal is real, but the tool is wrong. This creates the definitive "Call to Action," confirming that **Feature Engineering** is the mandatory bridge required to unlock the data's potential.

4.1.3 ACT 3: THE RESOLUTION

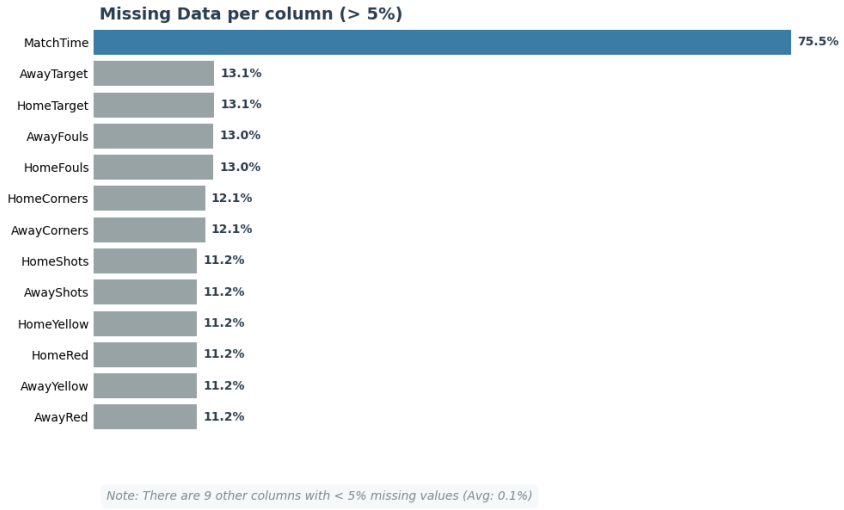
To provide a cohesive conclusion to our narrative strategy, we employ the "Tie-Back" technique, returning directly to the fundamental question posed in Act I: *Is the "Cinderella Effect" merely random noise, or is it a latent pattern waiting to be found?* Throughout the narrative arc, we guided the audience from the initial skepticism of the Setup, through the proven "Illusion of Raw Data" in the Conflict, to the definitive empirical proof in the Resolution. We have successfully demonstrated that these high-variance events are not products of stochastic chaos, but are the deterministic results of specific competitive dynamics—specifically, the interaction between an improving underdog and a stagnant favorite.

With this new understanding established, we issue a clear **Call to Action** for future predictive modeling. Analysts must abandon the reliance on static, absolute metrics like raw Elo ratings or simple goal counts, as our analysis proves they lack the discriminative power to handle complexity. Instead, the analytical framework must fundamentally shift toward **Differential Features**. The predictive signal resides effectively in the *gap* between competitors—represented by variables like `Elo_diff` and `Momentum_diff`—rather than in the isolated attributes of the teams themselves.

Furthermore, models must explicitly incorporate dynamic vectors such as **Fatigue** and **Momentum** to capture the trajectory of performance, rather than just historical averages. By adopting this strategic shift, we transition from a methodology that merely guesses based on reputation to one that detects based on performance reality. The signal for rare events exists; rigorous **Feature Engineering** provides the necessary lens to observe it.

4.2 Visualization Analysis

4.2.1 CHART 01: Missing values per column



Percentage of Missing Values per Columns

a. Reason for Choosing the Chart Type

The horizontal bar chart is the definitive choice for comparing nominal variables with high cardinality and extended text labels. Unlike vertical column charts, which would necessitate label rotation or abbreviation—thereby increasing cognitive friction—the horizontal orientation preserves the legibility of specific column names (e.g., `HomeCorners`) alongside their quantitative values. The use of length encoding provides the highest degree of perceptual accuracy for ratio comparisons, allowing the viewer to instantly grasp the massive magnitude discrepancy between `MatchTime` and the secondary variables. Furthermore, the descending sort order effectively organizes the data by severity, leveraging a Pareto-style structure to force immediate focus on the primary data quality bottleneck before scanning the "long tail" of minor errors.

b. Preattentive Attributes & Visual Hierarchy

The visualization utilizes a "pop-out" effect through selective color saturation, rendering the `MatchTime` bar in a deep, authoritative blue while relegating the remaining categories to a recessive, uniform gray. This chromatic contrast exploits the brain's preattentive processing to separate the critical signal (the 75.5% outlier) from the contextual noise before the viewer consciously reads the text. Hierarchy is further reinforced by the top-left positioning of the title and the primary bar, aligning with the Z-pattern of Western reading behavior to ensure the most critical information is

seen first. Typography plays a supporting role; the bolding of the percentage values adds visual weight to the data points, ensuring the specific metrics command more attention than the categorical labels themselves.

c. Clutter Reduction & Information Efficiency

Clutter is aggressively minimized through the elimination of the x-axis, tick marks, and vertical gridlines, removing all non-data ink that does not convey information. By employing direct labeling (placing the **75.5%** and other values immediately adjacent to the bars), the design utilizes the Gestalt principle of proximity, eradicating the need for the eye to scan back and forth between the data and a distant reference axis. Information efficiency is further refined by the editorial decision to filter out variables with less than 5% missing data; rather than visualizing negligible noise, this information is compressed into a subtle summary footnote. This constraint ensures the chart remains a decision-support tool rather than a comprehensive data dump, maximizing the signal-to-noise ratio for the viewer.

4.2.2 CHART 2: Elo Distribution

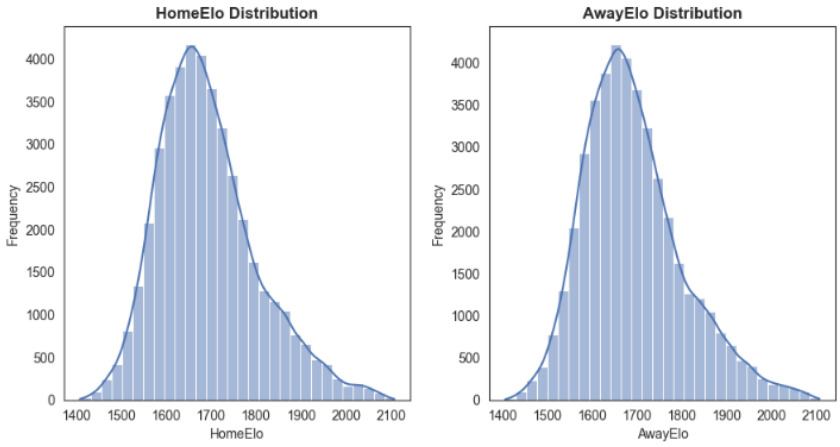


Figure 3.3: Elo Distribution

a. Reason for Choosing the Chart Type

The combination of histograms with overlaid Kernel Density Estimation (KDE) curves serves as the definitive method for analyzing the underlying frequency distribution of continuous numerical data. By employing a side-by-side "small multiples" layout, the visualization allows for the simultaneous assessment of central tendency, variance, and skewness across two distinct populations (`HomeElo` vs. `AwayElo`) without the visual occlusion common in overlapping charts. This structure effectively reveals the non-normal, right-skewed nature of the data, highlighting that while

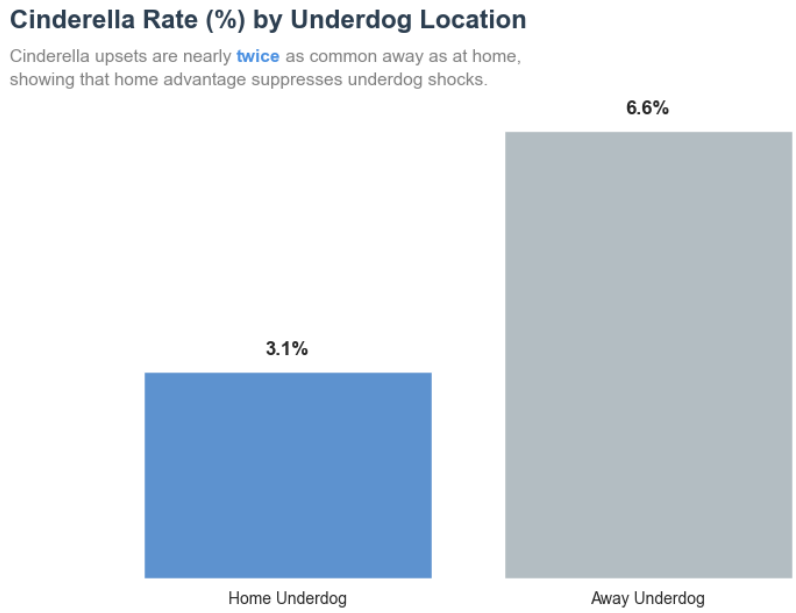
the majority of entities cluster around the 1650 baseline, a significant "long tail" of high-performance outliers exists. The dual representation—bars for granular frequency and lines for continuous probability trend—provides a robust validation of the dataset's structural symmetry.

b. Preattentive Attributes & Visual Hierarchy

Visual hierarchy is established through the contrast between the volumetric weight of the light-blue bins and the sharp definition of the dark-blue KDE line. The darker curve acts as the primary attentional anchor, guiding the eye smoothly across the distribution's shape and effectively filtering out the jagged noise of individual bin variances. The identical geometric peaks in both panels utilize height and mass as preattentive cues, forcing immediate recognition of the mode as the dominant value. Furthermore, the uniform color consistency across both charts creates a semantic link, signaling to the viewer that these variables share the same unit of measure and scale, facilitating rapid comparative processing.

c. Clutter Reduction & Information Efficiency

Information efficiency is maximized by synchronizing the x and y-axis scales (1400–2100 range and 0–4000 frequency) across both panels, eliminating the cognitive load required to mentally rescale unmatched axes. The absence of background gridlines and fill patterns ensures a high signal-to-noise ratio, allowing the distribution shape to stand out clearly against the white negative space. Text is kept minimal, with labels restricted to essential axis descriptors and titles, avoiding the redundancy of annotating individual bar heights. This minimalist approach forces focus onto the macro-level pattern—the mirrored distribution shapes—rather than micro-level data points.



Cinderella Rate by Underdog Location

a. Reason for Choosing the Chart Type

A vertical column chart is selected to execute a sharp binary comparison, using length encoding to physicalize the ratio between "Home" and "Away" underdog performance. This simple structure is ideal for validating the specific narrative claimed in the subheading: that away upsets occur nearly twice as often. By placing the bars side-by-side on a shared baseline, the visualization eliminates the need for complex mental modeling, allowing the magnitude of the disparity to serve as the primary visual evidence supporting the chart's declarative statement.

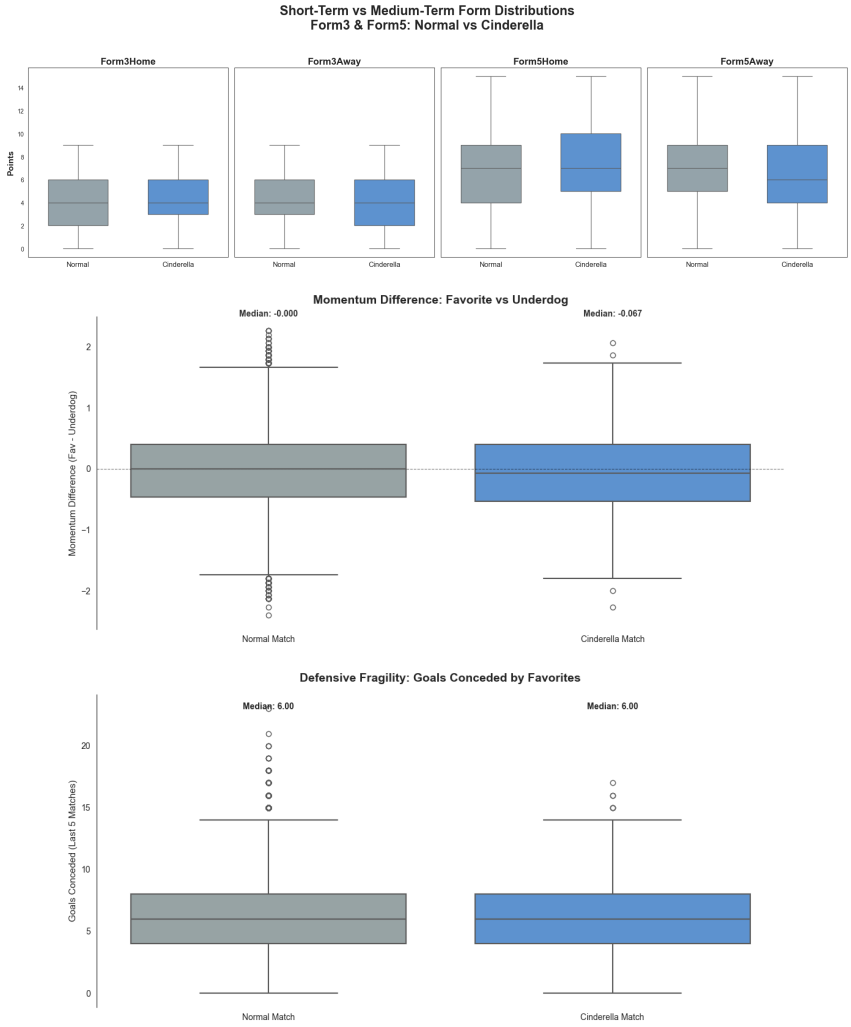
b. Preattentive Attributes & Visual Hierarchy

The visual hierarchy is anchored by the **left-aligned title and explanatory subheading**, which utilize the natural Z-pattern reading path to explicitly state the core insight ("nearly twice as common") before the viewer even processes the data. This "text-first" strategy primes the audience with the conclusion, reducing the cognitive work required to interpret the bars. Attention is then directed to the "Home Underdog" bar through a distinct blue hue, which creates a sharp contrast against the neutral gray of the "Away" category. This selective coloring creates a focal point, visually connecting the data back to the "suppression" narrative mentioned in the text.

c. Clutter Reduction & Information Efficiency

Information efficiency is driven by the minimalist removal of the y-axis, gridlines, and tick marks, effectively treating the chart area as a clean canvas. By integrating bold

percentage labels directly at the top of each bar, the design employs the Gestalt principle of proximity, allowing for instant value retrieval without the eye-travel required by traditional axes. This removal of non-data ink maximizes the signal-to-noise ratio, ensuring that the viewer's focus remains locked on the narrative flow—from the insight-rich subheading down to the contrasting bars—without distraction from decorative scaffolding.



Boxplot

a. Reason for Choosing the Chart Type

A unified series of comparative boxplots is selected to rigorously audit the statistical distributions of pre-match metrics (`Form`, `Momentum`, and `Defensive Fragility`) across "Normal" and "Cinderella" outcomes. This format supersedes simple aggregation, as it exposes the full interquartile ranges (IQR), whisker extents, and outliers, revealing the structural identity between the two datasets rather than just central

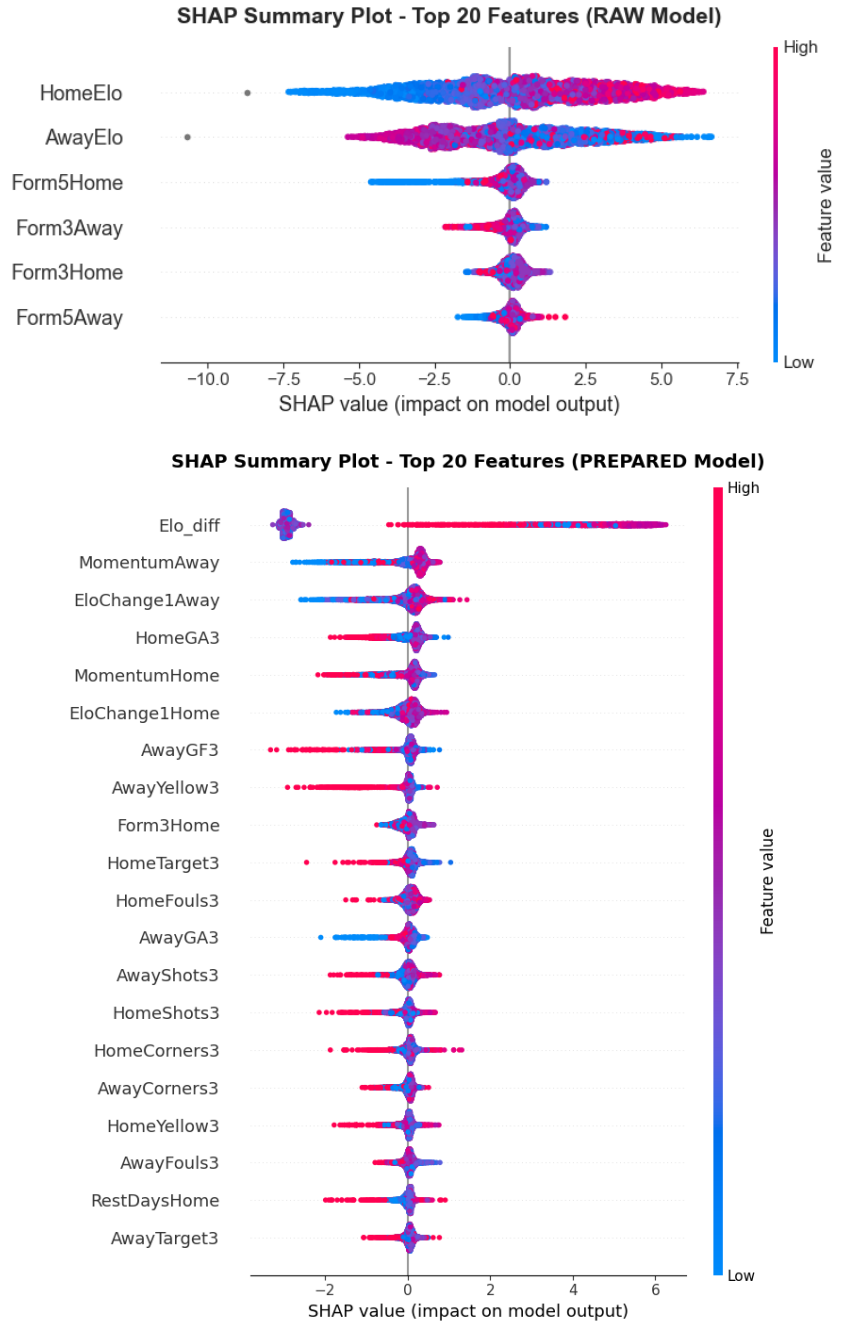
tendencies. The visualization validates a "null result" insight: the striking alignment of the distributions demonstrates that neither form fluctuation, momentum differentials, nor defensive fragility serve as reliable discriminators for upset victories. The consistent boxplot architecture allows for rapid, cross-metric validation of this stability without altering the interpretative framework.

b. Preattentive Attributes & Visual Hierarchy

Visual hierarchy is reinforced through a persistent chromatic schema, where saturated blue isolates the "Cinderella" variable against the neutral gray control group across all three panels. This hue persistence creates a cohesive visual grammar, allowing the viewer to instantly track the variable of interest across distinct analytical dimensions. In the Momentum and Fragility charts, explicit numeric annotations (e.g., "Median: -0.000", "Median: 6.00") function as high-priority cognitive anchors, reinforcing the geometric symmetry of the medians. Outlier markers are rendered as hollow circles to indicate extreme variance without overpowering the central tendency data, effectively separating the signal of the mass distribution from the noise of edge cases.

c. Clutter Reduction & Information Efficiency

Information efficiency is maximized by stripping away non-data ink, specifically horizontal gridlines and background shading, which would otherwise obscure the subtle alignment of the medians. The faceted "Short-Term vs Medium-Term" panel utilizes shared y-axes to eliminate redundant labeling, while the "Momentum" chart employs a single, low-weight dotted baseline to establish the zero-point context without visual heaviness. By restricting annotations to essential median values and axis descriptors, the design maintains a high signal-to-noise ratio, forcing the focus exclusively onto the overlapping physical dimensions of the boxes to confirm the statistical similarity between the groups.



SHAP

a. Reason for Choosing the Chart Type

The SHAP summary beeswarm plot is selected to deconstruct the "black-box" predictive model, simultaneously quantifying global feature importance and visualizing local interaction effects. This structure **ranks features by descending importance from top to bottom** based on mean absolute SHAP values, ensuring the primary predictive drivers are identified immediately. Unlike traditional feature im-

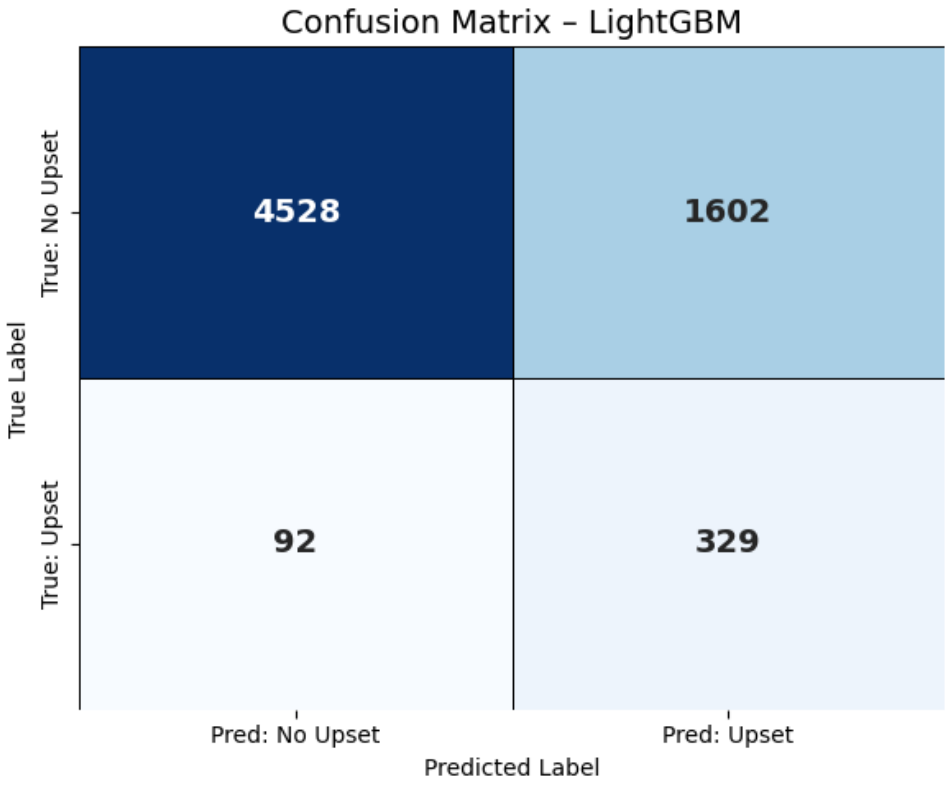
portance bar charts that aggregate data into a single scalar, the beeswarm format preserves directional impact and distributional spread, revealing complex non-linear relationships between input values and output predictions without obscuring individual data points.

b. Preattentive Attributes & Visual Hierarchy

Visual hierarchy is rigorously established through vertical positioning and spectral color coding. The most critical variables are anchored at the top to command primary attention, while the hue contrast between red (high feature value) and blue (low feature value) functions as a dominant preattentive cue. This color mapping allows the brain to instantly decode correlation directionality—for instance, red dots shifting to the negative side signal an inverse relationship. Horizontal positional encoding defines the magnitude of impact, sharply differentiating dense clusters near the zero-baseline from outliers with extreme influence on the model's output.

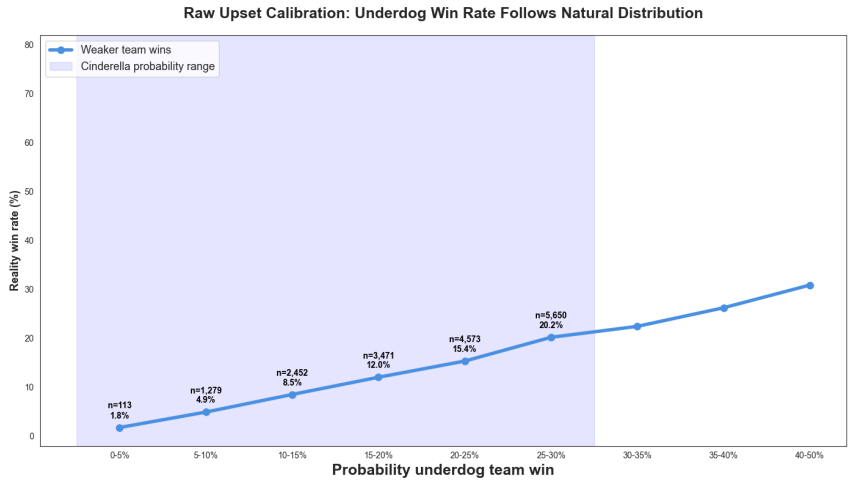
c. Clutter Reduction & Information Efficiency

Information efficiency is optimized by eliminating background gridlines and container borders, directing focus entirely toward the density and spread of the individual data points. This design consolidates the utility of multiple scatterplots into a single high-density view, achieving a superior data-ink ratio. A minimalist vertical color bar replaces complex legends, while the y-axis is stripped of tick marks to reduce visual noise. This reductionist approach minimizes cognitive load, ensuring the viewer focuses on the distribution patterns of the SHAP values rather than decorative scaffolding.



Confusion Matrix of LightGBM Model with Prepared Data

A heatmap-style confusion matrix is the optimal visualization for evaluating the granular performance of the LightGBM classifier beyond top-line accuracy metrics. By mapping cell density to color saturation, this format instantly exposes the specific nature of prediction errors, distinguishing clearly between false positives and false negatives. The matrix structure is essential for revealing the model's severe bias toward the majority class ("No Upset"), allowing for a precise diagnostic of recall versus precision trade-offs that a simple bar chart or scalar metric would obscure.



Raw upset calibration

a. Reason for Choosing the Chart Type

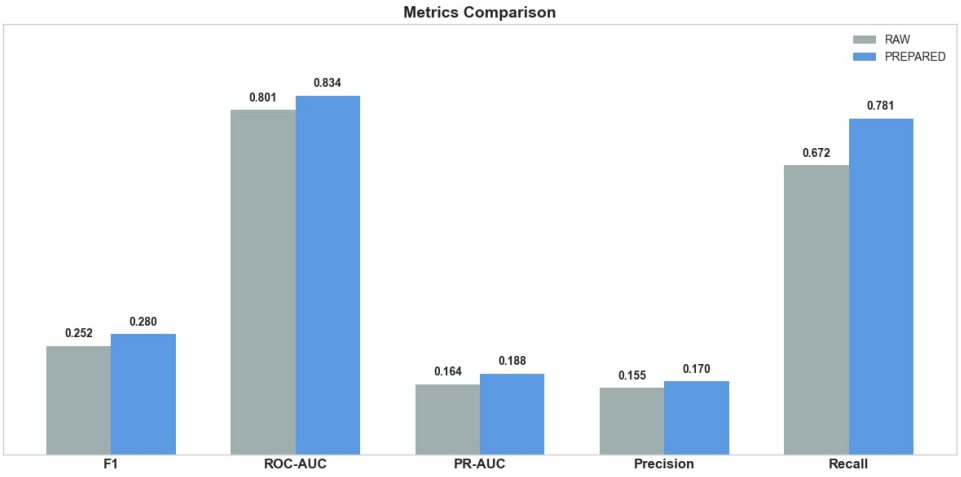
A calibration line chart is selected to visualize the correlation between predicted underdog probabilities and realized win rates, serving as a diagnostic tool for data distribution. By plotting aggregated bins connected by a continuous line, the visualization exposes the linear "natural distribution" of the dataset, confirming that raw upset rates scale predictably without structural deviation. This format is critical for demonstrating the "tail slicing" phenomenon: it proves that Cinderella events (low probability outcomes) do not exhibit a distinct anomaly in raw data but simply occupy the extreme lower tail of a standard distribution. The chart establishes the analytical baseline that raw data alone mimics randomness and is insufficient for discrimination, validating the requirement for advanced preprocessing to reveal hidden signals.

b. Preattentive Attributes & Visual Hierarchy

Visual hierarchy is anchored by the large, translucent purple zone representing the "Cinderella probability range" (<30%). This area annotation utilizes color contrast to physically demarcate the "extreme tail" region from the rest of the distribution, immediately orienting the viewer to the specific subset under analysis. The solid blue trend line cuts through this region with uninterrupted linearity, serving as a preattentive cue of continuity; this signals that these rare events follow the standard statistical slope rather than forming a unique, separable cluster. Text annotations at each node reinforce this hierarchy by pairing low sample sizes (n-values) with win rates, visually quantifying the scarcity of events within the highlighted tail compared to the higher-probability buckets.

c. Clutter Reduction & Information Efficiency

Information efficiency is achieved through the aggregation of continuous probability data into discrete bins (e.g., 0–5%, 5–10%), preventing the visual noise of a raw scatterplot while preserving trend fidelity. The decision to embed detailed statistics (n-count and percentage) directly at the data nodes eliminates the need for cross-referencing a separate data table, keeping the statistical evidence proximal to the visual trend. Background gridlines are minimized, and axis ranges are tightly cropped to the relevant data window. This decluttered approach ensures that the primary insight—that Cinderella events are merely a linear continuation of normal underdog mechanics rather than a distinct pattern—remains the undisputed focal point.



Raw vs Prepared Model - Metrics Comparison

a. Reason for Choosing the Chart Type

A grouped bar chart is utilized to execute a direct, side-by-side benchmarking analysis between the "RAW" and "PREPARED" model states across multiple disjoint performance metrics. This spatial arrangement places the baseline (Raw) and the optimized version (Prepared) in immediate proximity, forcing a direct comparison of magnitude for every specific metric category. Unlike stacked bars, which obscure individual values, the grouped structure prioritizes the visualization of the "uplift" or "delta"—allowing the viewer to instantly validate that the data preparation stage yielded universal improvements, particularly in the critical Recall dimension.

b. Preattentive Attributes & Visual Hierarchy

Visual hierarchy is dictated by a strategic chromatic contrast: the "RAW" baseline is rendered in a recessive neutral gray, while the "PREPARED" model is encoded in a saturated, active blue. This color choice leverages preattentive processing to designate the gray bars as context and the blue bars as the signal, guiding the

eye to perceive the blue height superiority as the primary narrative. The specific arrangement—placing the blue bar to the right of the gray—mimics the natural left-to-right progression of time (Before → After), subconsciously reinforcing the concept of progression. The dramatic height difference in the "Recall" pair (0.672 vs 0.781) acts as a visual anchor, drawing immediate attention to the area of most significant impact.

c. Clutter Reduction & Information Efficiency

Information efficiency is maximized through the total elimination of the Y-axis, tick marks, and horizontal gridlines. By integrating precise numerical values (e.g., 0.834, 0.781) directly atop the bars, the design utilizes the Gestalt principle of proximity, allowing for instant value retrieval without the cognitive load of cross-referencing a distant scale. The X-axis spine is kept minimal, serving only to group the categories. This high data-ink ratio ensures that the viewer's attention is not diluted by chart scaffolding, focusing entirely on the comparative height differentials that prove the efficacy of the model preparation process.