Data Drift Analysis in Mixed Martial Arts Outcome Prediction: Enhancing Existing Models with Betting Odds and Online-Attention Signals

Cédric Lam

Master of Science Candidate

HEC Montréal

August 14, 2025

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Mixed Martial Arts (MMA) prediction has traditionally relied on fighter statistics derived from historical performance data. While effective, these *existing approaches* face two key limitations: (i) they capture only past performance without incorporating real-time market intelligence or public sentiment; and (ii) they struggle with concept drift as tactics, regulations, and athlete populations evolve (Gama et al., 2014; Widmer & Kubat, 1996).

This thesis addresses these limitations by **enhancing existing MMA prediction models** through the systematic incorporation of two novel data sources: **betting odds** (capturing market wisdom) and **online attention signals** (capturing public sentiment via Google Trends). Rather than replacing traditional fighter statistics, we demonstrate how these complementary signals can significantly improve predictive performance when integrated into existing modeling frameworks.

We validate the enhancement approach using **event-based walk-forward validation** that simulates real deployment. This design retrains on the data immediately preceding each UFC event and evaluates only that card, preventing future leakage by construction while yielding a time-ordered series of estimates for accuracy and calibration.

The enhancement framework is applied to two complementary model architectures: a **binary classifier** for win/loss prediction and a **multiclass model** that predicts six mutually exclusive outcomes incorporating the method of victory. Both models

integrate the enhanced feature set combining traditional fighter statistics with market intelligence and online attention signals.

A key methodological contribution of this thesis is demonstrating how to effectively integrate diverse data sources while maintaining interpretability. The feature space combines traditional fighter metrics with market-derived features (odds differentials, implied probabilities) and public attention indicators (Google Trends search volumes), all engineered as interpretable *ratios, differences, and simple aggregates* combined with SHAP (Lundberg & Lee, 2017) for model explanation. Through systematic evaluation, this thesis demonstrates that enhancing existing models with betting odds and online attention signals provides substantial performance improvements: betting odds contribute approximately 14.4% and Google Trends contribute 8–10% of total model importance (as demonstrated in Chapter 5), validating the enhancement approach while building upon the foundation of sports betting market efficiency research (Sauer, 1998; Vaughan Williams, 1999).

1.2 Research Objectives and Questions

This thesis focuses on enhancing existing MMA prediction models through the systematic integration of market intelligence (betting odds) and public sentiment (online attention signals), with rigorous evaluation emphasizing calibration and interpretability.

O1 — Enhanced pre-fight models. Develop and benchmark enhanced pre-fight classifiers that integrate traditional fighter statistics with betting odds and Google Trends features, using event-based walk-forward validation while optimizing both *Accuracy* and *Brier* (Brier, 1950) to study the accuracy vs calibration trade-off in enhanced models.

O2 — Dual enhancement comparison. Compare the effectiveness of the enhancement approach across two complementary prediction tasks: binary win/loss prediction and multiclass outcome prediction (Decision, KO/TKO, Submission $\times$ fighter). Both enhanced models integrate identical feature sets combining traditional statistics with market and attention signals, ensuring fair comparison of the enhancement approach's effectiveness.

O3 — Economic validation of enhancements. Quantify whether the enhanced models' superior calibration translates into superior economic value through betting applications. Evaluate how the integration of betting odds and online attention signals improves practical performance compared with traditional statistics-only baselines.

O4 — Enhancement stability under drift. Assess the temporal stability of the enhancement features (betting odds and Google Trends) compared with traditional fighter statistics. Apply drift detection methods including KS (Massey, 1951) and PSI (Siddiqi, 2017) tests to evaluate whether market-based and attention-based enhancements provide more robust signals than traditional features when facing concept drift, particularly during disruptions like COVID-19.

These objectives lead to the following research questions:

RQ1 How effectively do betting odds and online attention signals enhance traditional MMA prediction models, and which objective (Accuracy vs. Brier) best captures the performance improvements from these enhancements?

RQ2 How do the enhancements (betting odds + Google Trends) contribute to predictive performance across different model architectures (binary vs. multiclass), and what are their relative feature importance patterns?

RQ3 Do the enhanced models' superior calibration translate into positive economic value, demonstrating that the integration of market intelligence and public sentiment provides practical benefits beyond traditional approaches?

RQ3b For betting applications, do enhanced binary models outperform enhanced multiclass models when collapsed to binary probabilities, and under what market conditions do the enhancements provide the greatest advantage?

RQ4 Are the enhancement features (betting odds and Google Trends) more stable than traditional fighter statistics when facing concept drift, particularly during external disruptions like COVID-19?

1.3 Thesis Roadmap

This thesis presents a systematic journey from understanding the current state of MMA prediction to demonstrating how market intelligence and public sentiment can significantly enhance traditional approaches. Each chapter builds upon the previous one, creating a comprehensive framework for enhanced MMA prediction.

Chapter 2: Literature Review establishes the theoretical foundation by surveying existing MMA prediction approaches and their limitations. This chapter is crucial for understanding why traditional fighter statistics alone are insufficient in the face of concept drift and evolving fighter strategies. We explore the rich literature on betting market efficiency and online attention signals, providing the theoretical justification for why these novel data sources should enhance prediction accuracy. This groundwork directly motivates our enhancement approach and sets the stage for our methodological contributions.

Building on this foundation, **Chapter 3: Data Collection & Feature Engineering** tackles the practical challenge of integrating diverse data sources. This chapter matters because the success of any machine learning approach depends critically on feature quality. We detail how to transform raw betting odds into calibrated probability features and convert Google Trends data into meaningful attention signals. The careful engineering of these features—as interpretable ratios, differences, and aggregates—ensures that our enhanced models remain explainable while capturing complex market dynamics and public sentiment patterns.

Chapter 4: Experimental Framework then presents our unified evaluation methodology, which is essential for fair comparison across different modeling approaches. The event-based walk-forward validation protocol developed here prevents the data leakage that plagued earlier MMA prediction studies while mimicking real-world deployment conditions. This chapter establishes the rigorous experimental standards that allow us to make confident claims about the value of our enhancements.

With the methodological groundwork established, **Chapter 5: Binary Classification Results** demonstrates the first major payoff of our enhancement approach. This chapter reveals how betting odds and Google Trends signals dramatically improve basic win/loss prediction, with XGBoost models optimized for calibration (Brier score) achieving the best performance. The results here validate our core hypothe-

sis that market intelligence and public sentiment contain predictive signal beyond traditional fighter statistics.

Chapter 6: Multiclass Classification extends these findings to the more challenging task of predicting not just who wins, but how they win. This chapter is vital for demonstrating the generalizability of our enhancement approach—showing that betting odds and Google Trends improve predictions across different model architectures and prediction tasks. The multiclass results reveal interesting patterns about which fight outcomes (Decisions vs. KO/TKO vs. Submissions) benefit most from each enhancement type.

A critical question for any predictive system is stability over time, which **Chapter 7: Temporal Stability Analysis** addresses through comprehensive drift detection. This chapter proves essential by showing that our enhancement features (particularly betting odds) maintain stability better than traditional fighter statistics during major disruptions like COVID-19. This finding has profound implications for model deployment and maintenance in production environments.

Chapter 8: Economic Validation provides the ultimate test of our enhancements by translating improved calibration into betting returns. This chapter matters because it demonstrates real-world value—showing that better-calibrated predictions from our enhanced models generate positive returns even when betting into efficient markets. The economic validation proves that our improvements are not just statistically significant but practically meaningful.

Finally, **Chapter 9: Conclusion** synthesizes all findings to provide actionable guidance for practitioners. This chapter distills our contributions into clear recommendations for integrating market intelligence and public sentiment into existing MMA prediction frameworks, while acknowledging limitations and charting promising directions for future research. The conclusion ensures that our theoretical and empirical contributions translate into practical impact for the MMA analytics community.

Chapter 2

Literature Review

2.1 Scope and Structure of the Review

The convergence of machine learning, sports analytics, and real-time data streams has created unprecedented opportunities for advancing predictive modeling in combat sports. Mixed Martial Arts (MMA), with its multi-dimensional skill requirements and dynamic competitive landscape, presents unique challenges that amplify the complexities inherent in sports prediction. Since the Ultimate Fighting Championship's (UFC) inception in 1993, the sport has undergone continuous evolution—from rule modifications and scoring system overhauls to tactical innovations and fighter development paradigms. These changes manifest as concept drift, where the statistical relationships between predictors and outcomes shift over time, potentially degrading model performance.

Current MMA prediction models achieve accuracy rates ranging from 65.52% (Wang & Zhang, 2024) to 80% (Berthet, 2023), with most studies reporting performance in the upper 60s to low 70s range. Despite these advances, they largely neglect the temporal stability of their predictions. The literature reveals three critical gaps. First, systematic drift detection mechanisms tailored to combat sports remain absent. Second, the rich information contained in betting markets is rarely integrated with technical performance data. Third, social media signals, despite representing real-time fan engagement and potential insider knowledge, have not been systematically exploited for MMA outcome prediction.

This review synthesizes five interconnected research streams to address these gaps.

First, we examine concept drift theory and detection methodologies applicable to sports contexts, establishing the theoretical foundation for understanding how and why predictive relationships change over time. Second, we analyze current machine learning approaches in MMA prediction, identifying their strengths, limitations, and the ceiling effects encountered. Third, we explore betting market efficiency and the integration of odds data into predictive models, viewing odds not as gambling tools but as aggregated expert knowledge. Fourth, we investigate social media signal extraction for sports forecasting, focusing on information value rather than sentiment analysis. Finally, we consider adaptive learning techniques for maintaining model performance over time, emphasizing practical implementation in production systems. Through this synthesis, we establish the theoretical and empirical foundations for developing drift-aware MMA prediction systems that leverage multiple information sources.

2.2 Concept Drift

This section establishes the theoretical foundation for the thesis by reviewing the phenomenon of *concept drift*, where the statistical properties of the data-generating process change over time. In dynamic domains like sports, recognizing and handling concept drift is critical for building robust prediction models. Concept drift refers to changes in the underlying distribution of features or outcomes that invalidate patterns learned from past data (Lu et al., 2019). In other words, the relationships that held true for predicting fight outcomes in the past may shift as the sport evolves—fighters adapt their styles, new training techniques emerge, and organizational changes (e.g., rules or matchmaking policies) alter the competition. If such drift is not addressed, model performance can degrade substantially because patterns learned on outdated data no longer apply to the new environment (Bayram et al., 2022; Lu et al., 2019). Consequently, concept drift has been identified as a root cause of deteriorating predictive accuracy in many data-driven systems when the environment changes.

Concept drift can manifest in different forms: it may be *gradual*, such as the slow evolution of fighters' skills or strategies over years; *sudden*, following abrupt rule changes or a breakout performance that shifts perceptions; or *recurring*, through cyclical patterns that come and go. In MMA, one can hypothesize gradual drift as

the sport's metagame changes (for instance, the rise of a particular fighting style) and sudden drift in the aftermath of major upsets or championship fights that redefine competitive dynamics. Indeed, recent complex network analysis of UFC matchups found evidence that the structure of competition has evolved significantly from the early years to the present (Castillo et al., 2025), indicating that the context in which predictions are made is not stationary. It is therefore prudent to expect that a model trained on historical fight data will face a shifting target over time.

To ensure valid evaluation under concept drift, one must use time-aware validation strategies. Traditional k-fold cross-validation, which assumes i.i.d. data shuffling, can be misleading in the presence of temporal dependencies or evolving data (Bergmeir & Benítez, 2012). Instead, this thesis adopts an *event-based walk-forward validation*, where the model is trained on past fights and tested on the next chronological event, then updated sequentially. This approach mirrors how a model would be deployed in real time, always training on past data and predicting future events in order, thus respecting the temporal order of fights. Such walk-forward validation is recommended for non-stationary problems to avoid lookahead bias and to accurately measure how performance changes as new data arrives. In our context, this means iteratively retraining or updating the model as each new fight outcome becomes available and evaluating performance on the next fight (or batch of fights). By examining performance epoch by epoch, we can also diagnose drift; for example, if prediction error spikes after a certain date or event, that signals a possible concept drift event requiring adaptation.

This thesis also explicitly investigates *post-fight drift* by analyzing how model calibration and accuracy evolve after each event. This approach determines if the incorporation of a new result significantly shifts the underlying patterns. Concept drift research emphasizes monitoring for changes over time (Hinder et al., 2024), and our work contributes to this by examining on a fight-by-fight basis whether the predictive model's errors indicate a shift. For instance, a string of unexpected upsets might suggest the prior model was out-of-sync with the true odds. By combining sequential validation with post-event performance analysis, our methodology aligns with best practices for handling evolving data streams in machine learning (Hinder et al., 2024; Lu et al., 2019). This focus on the temporal nature of the data lays the groundwork for selecting appropriate machine learning models, which are discussed next.

2.3 Machine Learning for MMA Outcome Prediction

Having established the challenge of concept drift, this section surveys the application of machine learning (ML) models to MMA outcome prediction, identifying common approaches and their limitations in a dynamic environment. ML techniques have become integral to predicting sports outcomes, including MMA fight results. Early applications to MMA focused on binary classification of fight winners using historical data. For example, Hitkul et al. (2019) performed a comparative study of algorithms for UFC fight prediction, applying methods like logistic regression, support vector machines, and decision trees. Such studies demonstrated the feasibility of data-driven predictions, albeit often on limited feature sets (e.g., win-loss records, physical attributes) and with moderate accuracy in the range of 60–65%.

More recent work has significantly expanded both data and methodology. Yan Sheng et al. (2024), for instance, leverage machine learning not only to predict match outcomes but also to identify influential factors that impact fight results, offering strategic insights to fighters and coaches. Their approach underscores that beyond pure prediction, ML models can reveal which attributes—such as reach, striking accuracy, or ground game—are most predictive of success, aligning with the growing interest in interpretability. Similarly, Wang and Zhang (2024) incorporate high-level features like fighter style archetypes to improve predictive performance, recognizing that stylistic matchup information can be crucial. In parallel, novel modeling approaches have appeared. A notable example is the use of a Markov chain model for MMA by Holmes et al. (2023), who simulate contests on a per-round basis rather than treating outcome prediction as a black-box classification. Their model achieved well-calibrated probability forecasts by explicitly modeling the progress of a fight, indicating an alternative to standard ML classification that can naturally account for the method-of-victory and time dynamics. On the pure ML side, state-of-the-art models often use ensemble methods like gradient boosted trees or random forests due to their strong performance on structured sports data. Indeed, some recent MMA prediction systems have reported high accuracy; for example, Berthet (2023) report approximately 80% accuracy when using in-fight live statistics to predict fight outcomes in real time. This illustrates the upper bound of performance when rich data

are available during the bout. In a pre-fight setting, where only historical and pre-match features are known, the achievable accuracy is typically lower, but ML models can still outperform simplistic baselines.

A distinguishing feature of this thesis is framing fight prediction as a *multiclass classification* problem to predict not only the winner but also the method of victory (e.g., knockout, submission, decision). While traditional studies and betting odds usually treat winner prediction as binary, the manner in which a fighter wins is of great interest and has its own odds markets. Our approach aligns with these "method of victory" predictions by creating a multiclass outcome space (e.g., Fighter A by KO, Fighter A by submission, etc.). This formulation provides a richer prediction output and allows for a more granular analysis of model confidence across different outcome types. Although less common, multiclass outcome prediction has been explored indirectly (e.g., Holmes et al. 2023) and has been documented in other sports like tennis and soccer, where it has revealed patterns invisible to binary approaches. By addressing the multiclass problem directly, we extend prior work and enable a more fine-grained evaluation of predictive performance.

Finally, this work emphasizes model interpretability. Complex ML models can be "black boxes," but understanding what drives a prediction is valuable for both scientific inquiry and stakeholder trust. We employ SHAP (SHapley Additive exPlanations) analysis (Lundberg & Lee, 2017) to interpret our fight outcome predictions. SHAP is a game-theoretic approach that assigns each feature an importance value for a particular prediction, helping to explain whether a fighter's recent win streak, reach advantage, or age contributed most to the predicted win probability. The use of SHAP in sports prediction is relatively novel but is gaining traction as practitioners seek to justify decisions. By incorporating SHAP, we follow a broader trend of integrating explainability into predictive modeling, aligning with recent UFC-focused research that calls for providing insights into *why* a model favors one fighter (Yan Sheng et al., 2024). This review underlines that combining predictive power with interpretability is a modern best practice. While model interpretability provides insight into prediction drivers, an equally important external signal comes from betting markets, which reflect the aggregated wisdom of the crowd. The next section explores how this information can be integrated.

2.4 Betting Odds and Prediction Markets

Having reviewed the internal mechanics and interpretability of ML models, we now turn to an external source of predictive information: the betting market. This section examines the role of betting odds as both a benchmark for model performance and a source of predictive features. Betting odds represent aggregated public and expert opinion about fight outcomes and thus serve as a valuable reference point. In a betting market, odds (when converted to implied probabilities) often encapsulate a wide array of information—fighter skill levels, training camp reports, and perhaps even insider knowledge—effectively making them an "oracle" baseline that can be difficult for models to outperform. In this thesis, odds are used (i) as predictive features and (ii) as an external yardstick against which model calibration is judged. Indeed, odds in sports like MMA are, in aggregate, usually reasonably well-calibrated predictors of winners.

However, it is well-documented that betting markets are not perfectly efficient. One known imperfection is the *favourite–longshot bias*, where outcomes with low probability (longshots) are overvalued by the odds relative to their true chances, and favourites are slightly undervalued (Berkowitz et al., 2016). In practice, this means betting odds tend to overstate the chances of underdogs winning, leading to systematically lower returns on longshot bets (Berkowitz et al., 2016). Such biases present an opportunity for a well-tuned model to exploit by identifying when a favourite is under-bet or an underdog is over-hyped. Machine learning models have been applied to sports betting with success in identifying these small inefficiencies. Hubáček et al. (2019), for example, demonstrated that an ML model could exploit odds in soccer betting markets to achieve positive returns by predicting outcomes more accurately than the market in certain situations. Their approach highlights that a predictive model's goal in a betting context is not just accuracy, but identifying value—cases where the model's estimated win probability differs significantly from the implied probability in the odds.

In evaluating such models, traditional accuracy is an insufficient metric; the calibration of predicted probabilities becomes crucial. A model may predict many winners correctly, but if its probability estimates are poorly calibrated (e.g., consistently overestimating underdog win chances), it may still lose money when used for betting decisions. Recent research by Walsh and Joshi (2024) directly addresses this, showing

that using calibration as a model selection criterion leads to greater betting returns than using accuracy alone. This finding reinforces that our thesis must carefully consider probability calibration. We therefore devote attention to calibrating our model's output probabilities to ensure that a predicted 70% win chance for a fighter materializes as a win approximately 70% of the time in the long run. Proper calibration ensures that the model's confidence can be trusted and directly compared to betting odds. Following examples from Wheatcroft (2020) and Hubáček et al. (2019), our thesis assesses not just predictive skill (e.g., accuracy, Brier score) but also whether the model could generate positive returns in a betting simulation. This dual evaluation aligns with scientific literature and practical expectations in sports analytics. The detailed results of this dual evaluation, with a specific focus on model calibration against market odds, are presented in Chapters 5 through 8. While betting odds encapsulate significant information, they may not capture all real-time dynamics, such as public hype. The subsequent section investigates another data source—public attention signals—to determine if they provide complementary predictive value.

2.5 Public Attention Signals

Beyond the structured data of fight statistics and betting odds, modern data streams offer novel signals reflecting public interest. This section reviews the literature on using these *public attention signals*, particularly from Google Trends, as a potential feature source for improving predictive models. Athlete performance can be influenced by factors beyond physical and statistical attributes, notably the level of public attention surrounding a contest. Our thesis explores these signals by focusing on Google Trends search volume data, which captures broad interest levels in near real-time. The use of Google Trends for prediction has grown dramatically in the last decade across fields including finance, politics, and sports. A comprehensive review by Jun et al. (2018) notes that its application has shifted from descriptive analytics to forecasting, indicating that Google Trends has matured into a tool for providing predictive signals.

In sports, search volume may proxy public sentiment, athlete marketability, or even insider expectations; for example, a sudden increase in searches for an underdog might indicate rumors or confidence in an upset. Within MMA, these signals have tangible

connections to the sport's dynamics. Castillo et al. (2025) found that Google search trends for fights correlate strongly with pay-per-view sales and audience engagement. While their work focuses on audience metrics, it substantiates that Google Trends data measures the attention a fight receives. For predictive modeling, one hypothesis is that a fighter with surging public interest might be performing well or has an X-factor not captured by past performance metrics. By incorporating Google Trends indices, such as the relative search popularity of each fighter, our model can test whether such metrics improve predictive power or calibration.

Previous literature has also explored social media as a source of public sentiment. Studies like Schumaker et al. (2016) and Wunderlich and Memmert (2021) used Twitter sentiment analysis to predict soccer match outcomes. However, real-time sentiment is less applicable for pre-fight prediction, and social media data for MMA can be challenging to collect and noisy. We therefore pivot to search-based attention signals, which are available historically and tend to be less sparse. While literature directly linking Google search data to MMA fight outcomes is sparse, making our work exploratory, the theoretical justification is that public attention can encapsulate unmodeled factors, from fighter popularity to training camp buzz. Our literature review suggests that harnessing such data has become an important trend in predictive analytics (Jun et al., 2018). The integration of diverse and dynamic data sources necessitates a modeling framework that can adapt over time. The final section of this review addresses this need by examining adaptive learning strategies designed to maintain model performance.

2.6 Adaptive Learning and Model Updating

The preceding sections have established the dynamic nature of the MMA prediction problem, highlighting concept drift and evolving data streams. This section synthesizes these threads by reviewing *adaptive learning* and model updating strategies, which are essential for maintaining predictive accuracy over time. Adaptive learning refers to the model's capability to update itself as new data becomes available, rather than being trained once and held static. The literature provides numerous strategies for adaptation, which fall broadly into two categories: passive adaptation, which involves continuous updating, and active adaptation, where the model updates only

when a drift detector signals a significant change (Hinder et al., 2024; Lu et al., 2019).

This thesis adopts a moderate, passive adaptive strategy: we retrain the model at regular intervals, specifically before each event in our walk-forward validation. This ensures the latest data is always included. The choice of retraining before each event is guided by our analysis of drift; if the relationship between features and outcomes is changing even subtly, incorporating the most recent fights should help the model adjust. Additionally, we explore adaptive calibration. Because probability calibration can also drift, we periodically recalibrate the model's output probabilities using the latest data, keeping its predictions probabilistically tuned as baseline win rates evolve. The need for this is supported by work like Walsh and Joshi (2024), who imply that continuous calibration evaluation is critical in betting scenarios.

While this thesis does not implement a standalone drift detector, our post-fight drift analysis serves a similar function. By monitoring rolling performance metrics, a sudden drop in accuracy or an uptick in Brier score would signal that the model's understanding has become misaligned, prompting more aggressive adaptation. This reflective approach ties into the concept of *performance-aware drift detectors* highlighted by Bayram et al. (2022). Fully adaptive systems, however, have trade-offs: adapting too frequently can lead to instability, while adapting too slowly misses the benefit. By structuring our evaluation as event-based sequential retraining, we inherently simulate an adaptive learning system that balances these concerns.

In conclusion, this literature review has traversed the key domains informing this thesis: the temporal challenge of concept drift, the application of machine learning to MMA, the informational value of betting markets and public attention, and the necessity of adaptive methods. The identified gaps and best practices from these fields collectively motivate the methodology of this thesis, which aims to construct a robust, multi-source, and drift-aware prediction system for MMA outcomes. The scientific foundation for this approach is strong, aligning with the framework of continuous learning under drifting concepts (Lu et al., 2019) and ensuring our predictive system remains as up-to-date as possible with the evolving state of the sport. The following chapters will detail the implementation and evaluation of this system.

Chapter 3

Data Collection, Exploration, and Feature Engineering

3.1 Overview

This section provides a comprehensive roadmap of the data engineering pipeline that forms the empirical foundation of our predictive system. The primary objective is to transform heterogeneous data sources into a unified, temporally consistent analytical framework that preserves the integrity of causal relationships while maximizing the information content available for modelling.

This chapter describes the full data pipeline that turns raw fight records, betting odds, and public-interest signals into a clean modelling table. The data are drawn from three complementary sources, each offering unique perspectives on fight outcomes. Table 3.1 presents these primary data sources, revealing significant variation in their temporal coverage and completeness. All code is orchestrated by the `DatasetBuilder` class introduced in §3.3, which ensures reproducibility and modularity in our data processing workflow.

Table 3.1: Primary data sources and coverage (as of 2025-07).

Source	Core contents	Rows	Cols	Years	Coverage
UFCStats	Events, fighter bios, fight-level statistics	10 793	393	1994–2025	100 % fights
OddsPortal	Closing money-line odds	4 478	6	2013–2025	46.7 % fights
Google Trends	Daily search interest	6 043	18	2006–2025	56.0 % fights

We observe from Table 3.1 that while UFCStats provides comprehensive historical coverage dating back to 1994, market-based indicators only became reliably available in the modern era—odds data begins in 2013, while Google Trends coverage starts in 2006. The substantial coverage gap between fight statistics (100%) and market-derived features (46.7% for odds, 56.0% for trends) necessitates careful treatment of missing data patterns. This heterogeneity motivates our use of explicit binary indicators (`_has_data` flags) rather than imputation for market features, preserving the information content inherent in data availability itself. This temporal structure has important implications for our modelling strategy, as discussed in Section 3.4.2.

The remainder of the chapter is organised to guide the reader through the complete data engineering journey. §3.2 details the acquisition layer, explaining how we systematically extract data from three distinct sources while respecting ethical scraping practices and rate limits. §3.3 then describes the modular processing architecture that enables parallel feature calculation and ensures reproducibility across different computational environments. §3.4 explains critical data validation and temporal integration procedures that prevent future data leakage—a fundamental requirement for valid backtesting and real-world deployment. §3.5 presents an extensive exploratory analysis that reveals key patterns in fight outcomes, validates our feature engineering choices, and identifies the most predictive signals. Finally, §3.6 enumerates the complete engineered feature space, documenting how raw inputs transform into 334 carefully crafted predictive variables.

This structured approach ensures that subsequent modelling efforts in Chapters 5 and 6 build upon a solid empirical foundation, with every data transformation justified and documented.

3.2 Data acquisition

This section describes the technical infrastructure and methodological choices underlying our data collection system. Each source presents unique challenges in terms of access patterns, rate limiting, and data structure, requiring carefully tailored acquisition strategies that balance comprehensiveness with responsible data practices.

3.2.1 Competitive performance: `UFCStats`

The foundation of our analytical framework rests on comprehensive fight statistics that capture the complete competitive history of mixed martial arts.

Fights, fighters, and events are scraped through a headless SELENIUM crawler that respects `robots.txt` and employs exponential back-off.

The scraping infrastructure produces three interconnected relational tables that collectively capture the multi-dimensional nature of MMA competition. First, the **events** table (1 209 rows, 5 columns) establishes the temporal and geographic context for each fight card, storing the event identifier, date, venue, city, and a completion flag that helps distinguish between fully recorded events and those cancelled or ongoing.

Second, the **fighters** table (3 899 rows, 7 columns) maintains a canonical registry of all athletes who have competed in major MMA organizations. Beyond basic identification through official names and nicknames, this table captures critical physical attributes: fighting stance (orthodox, southpaw, or switch), height in meters, reach in centimeters, and date of birth. These anthropometric measurements form the basis for numerous engineered features, as physical disparities between opponents often correlate with fighting style adaptations and outcome probabilities.

Third, the **fights** table (10 793 rows, 393 columns) represents our most granular data layer, containing exhaustive per-fighter metrics for every recorded bout. This rich dataset includes detailed breakdowns of significant strikes by target area (head, body, legs) and position (standing, clinch, ground), comprehensive grappling statistics (takedown attempts, success rates, submission attempts), control time measurements, and extensive contextual metadata such as title-bout status, referee assignment, and specific venue location. The breadth of these features—393 columns per fight—enables nuanced analysis of fighting patterns and outcome predictors that would be impossible with summary statistics alone.

Importantly, legacy promotions (WEC, Strikeforce, AFC) appear in the same schema, permitting longitudinal modelling that predates UFC debuts. This design decision reflects our recognition that many elite fighters developed their skills in these predecessor organizations, and excluding their pre-UFC records would create an incomplete picture of their capabilities and evolution.

3.2.2 Betting Markets: `OddsPortal`

Betting markets provide a unique window into collective wisdom about fight outcomes, aggregating information from thousands of informed participants including professional handicappers, statistical modelers, and domain experts. Our approach to odds acquisition prioritizes data quality and representativeness over sheer volume.

We parsed closing decimal odds for each fighter from over 30 bookmakers via static HTML requests. The decision to use closing odds rather than opening lines is deliberate and methodologically important: closing prices incorporate the maximum available information, including late-breaking news about injuries, weight-cutting difficulties, training camp reports, and shifts in public sentiment. After dropping duplicate sources that merely mirror other bookmakers' lines, we compute the median decimal line ω_i for fighter i in bout b. The median provides a robust central estimate that reduces the influence of outlier bookmakers who may have unusual exposure or house biases. This decimal odds value is then converted to its raw implied probability

$$\pi_i = \frac{1}{\omega_i}\,.$$

A critical consideration when working with betting data is the bookmaker's built-in profit margin. Because bookmakers build in a profit margin (also known as 'vigorish' or 'vig'), the raw implied probabilities for all outcomes in a bout sum to more than one. We therefore normalize by the total implied mass to obtain a true win probability that reflects the market's actual assessment

$$p_i \;=\; \frac{\pi_i}{\sum_{j=1}^{2} \pi_j}\,.$$

This normalization step serves two essential purposes: it enables fair comparison across different bookmakers with varying vig levels, and it provides proper probability estimates that sum to unity, as required for rigorous calibration assessment in our downstream analyses.

Odds data cover 3 615 of 7 745 eligible fights (46.7 %). This coverage limitation reflects the relatively recent emergence of liquid MMA betting markets, with comprehensive odds data becoming reliably available primarily after 2013. Despite this tem-

poral constraint, the subset with odds data proves highly informative: the favourite (the fighter with higher p_i) wins 50.9% of priced bouts. These results suggest minimal bias in simple favourite vs. underdog frequencies; paired with the low ECE (0.015), they indicate the market is reasonably calibrated for heavy favourites (see Section 3.5.8).

3.2.3 Public Attention: Google Trends

Daily search-interest indices were retrieved using the `pytrends` API (Jun et al., 2018). We query both fighters in each bout simultaneously to obtain a meaningful relative measure: Google Trends returns normalized indices (0-100) that are otherwise difficult to compare between single queries. The filter of the "MMA" category is consistently applied to disambiguate fighter names. To prevent fight day leakage, we retain an eight day window $[t-8, t-1]$ preceding each bout, explicitly excluding fight day searches from all analyses.

Public attention, as captured by Google Trends, has been shown to reflect changes in audience engagement and can serve as an early warning for feature drift in predictive models (Choi & Varian, 2012; Jun et al., 2018). Surges or declines in search interest often precede changes in betting markets or athlete visibility, making Trends an invaluable real-time covariate for outcome forecasting. The mechanism is intuitive: increased search activity may indicate breaking news (injury reports, training footage), viral moments (press conference incidents), or shifting public sentiment that has not been fully incorporated into betting prices.

Trends data were available for 6,043 bouts (56.0%). The coverage pattern exhibits interesting structure: better availability for recent fights and main-card matchups that generate substantial public interest, with preliminary bouts showing spottier coverage. Any missingness is handled systematically through binary indicator variables rather than imputation, preserving the information content inherent in data availability itself (see Section 3.4.2).

In summary, this section has described how three complementary data sources—comprehensive fight statistics from UFCStats, efficient market prices from OddsPortal, and dynamic public interest from Google Trends combine to provide a multi-faceted view of each matchup. No single source could deliver this richness in isolation; their synthesis enables predictive insights that emerge from the intersection of historical

performance, market wisdom, and public attention.

3.3 Processing architecture

This section details the modular processing infrastructure that transforms raw data into analysis-ready features. The architecture prioritizes reproducibility, computational efficiency, and maintainability through clear separation of concerns and parallel execution where possible.

Figure 3.1 illustrates the high-level data engineering flow that governs our processing pipeline. The pipeline follows a strict separation of concerns, with each stage handling a specific transformation responsibility. The central orchestrator `DatasetBuilder` coordinates the entire workflow: it loads sources from the acquisition layer, applies preprocessing transformations to ensure consistency, calculates features in parallel to maximize computational efficiency, and finally filters by temporal rules to maintain causal validity.

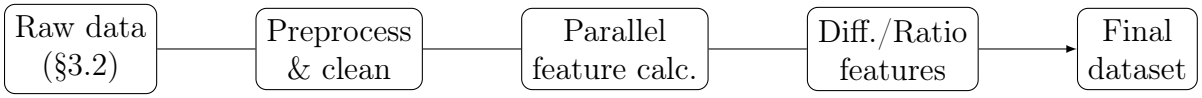


Figure 3.1: High-level data-engineering flow (simplified).

We observe in Figure 3.1 that the architecture enforces a unidirectional flow from raw data to the final dataset, preventing circular dependencies and ensuring reproducibility. Each processing stage can be independently tested and modified without affecting downstream components, a design choice that proved invaluable during iterative feature development. The parallel feature calculation stage represents a key optimization: by computing independent features concurrently, we reduce processing time from hours to minutes for the full dataset.

This modular architecture serves a dual purpose: it ensures that every data transformation is traceable and reproducible, while also enabling efficient computation through parallelization. The strict temporal filtering at the final stage guarantees that our features respect causality, a non-negotiable requirement for valid backtesting and live deployment.

3.4 Data cleaning and integration

This section describes the critical data quality procedures that ensure our dataset maintains both internal consistency and temporal validity. These cleaning steps address real-world data challenges including name variations, missing values, and the paramount concern of preventing future data leakage.

3.4.1 Temporal integrity

Preventing future data leakage stands as the most critical aspect of our data preparation pipeline. In backtesting scenarios, even subtle temporal violations can produce unrealistically optimistic performance estimates that fail catastrophically in live deployment.

All features are calculated using data strictly dated *before* the target bout. We implement a conservative approach: a helper function enforces a `cutoff` $= t - 1$ day rule, ensuring that even same-day information remains excluded. This one-day buffer accounts for timezone differences and late-reporting data that might otherwise contaminate our feature calculations. Listing 3.1 provides the pseudocode implementation of this temporal guard clause.

```
def ensure_temporal_integrity(dataset, fight_date):
    cutoff = fight_date - pd.Timedelta(days=1)
    return dataset[dataset['event_date'] < cutoff]
```

Listing 3.1: Guard clause preventing future leakage.

3.4.2 Missing-value strategy

Our approach to missing data reflects a key insight: the absence of information can itself be informative. We employ different strategies for different feature categories based on their missingness patterns and semantic meaning.

For **physical attributes**, we use statistical imputation methods that leverage the strong correlations within weight classes. Height values are imputed using weight class medians, as fighters within the same division tend to cluster around similar heights. Reach presents a more complex case—we employ linear regression using height and weight class as predictors, capitalizing on the biomechanical relationship

between arm span and stature. This approach preserves the natural variance in reach-to-height ratios while avoiding unrealistic values.

For **market features**, we take a fundamentally different approach. Rather than imputing odds or trends data, we preserve the NaN values and add explicit binary `_has_data` flags. This design choice reflects our recognition that missingness in market data is non-random: fights without odds data tend to be preliminary bouts or older contests from before liquid betting markets emerged. The `_has_odds` and `_has_trends` flags thus become features in their own right, allowing models to learn the systematic patterns in data availability.

Table 3.2 quantifies the missingness patterns across different feature blocks in our final dataset. We observe that odds and trends exhibit the highest missingness rates at 53.3% and 44.0% respectively, while deterministically calculated features achieve complete coverage. This pattern validates our decision to use explicit indicators rather than imputation for market-based features.

Table 3.2: Proportion of missing values by feature block (final dataset).

Feature block	Missing (%)	Handling rule
Physical	3.1	regression
Historical performance	0	deterministic aggregation
Odds (raw)	53.3	leave NaN, add flag
Trends (raw)	44.0	leave NaN, add flag
Engineered aggregates	0	derived from non-missing parents

The high correlation between odds and trends missingness suggests these features share common availability constraints—typically both are present for high-profile fights or both are absent for preliminary bouts. When examining the complete dataset before filtering, 40.5% of all fights (4,376 out of 10,793) lack both odds and Google Trends data, underscoring the substantial overlap in market feature unavailability. This structured missingness pattern is preserved through our `_has_data` flags, allowing models to exploit the information content in data availability itself.

3.4.3 Filtering inexperienced fighters

The quality of historical features depends critically on having sufficient past data from which to calculate meaningful aggregates. A fighter's first professional bout

provides no historical context, making accurate prediction nearly impossible without prior performance metrics.

To ensure that historical aggregates are meaningful, we implement a strict filtering criterion: fights where either athlete had zero prior professional bouts are discarded. This decision reflects a fundamental modeling constraint—without historical data, we cannot calculate performance trends, fighting style indicators, or career trajectories. The final analytic table therefore contains 7 745 contests, down from the initial 10 793 ($-28.2\,\%$). While this reduction appears substantial, it disproportionately removes debut fights from the early UFC era when record-keeping was less comprehensive, thereby improving overall data quality.

3.4.4 Target variable construction

The final step in our data preparation pipeline involves constructing the target variables that our models will predict. We implement a dual-target strategy that enables comprehensive evaluation of different modeling approaches.

The dataset supports two complementary target variables, enabling direct comparison of binary and multiclass modeling approaches using identical features. This parallel construction allows us to investigate whether predicting fight outcomes as a simple win/loss decision differs fundamentally from predicting the specific method of victory.

Binary target. The primary target `winner` $\in \{0, 1\}$ indicates whether fighter 1 wins (1) or loses (0). A critical consideration here is position bias: the raw UFCStats data always lists the winner first, creating a trivial 100/0 split that would lead to meaningless models. To ensure the model does not learn to favor the fighter listed first, we implement random position swapping: fighter positions are randomly exchanged in 50% of the dataset. This debiasing procedure results in a near-perfect balanced 49.7% / 50.3% win-loss split for the 'Fighter 1' position, eliminating any systematic advantage from list position.

Multiclass target. For richer prediction capturing both winner and finish method, we construct `fight_outcome` $\in \{0, ..., 5\}$. This six-class structure emerges naturally from the combination of two fighters and three primary finish methods (decision,

knockout/TKO, submission). Table 3.3 presents the complete label mapping:

Table 3.3: Multiclass label mapping for fight outcomes.

Label	Winner	Method	Frequency
0	Fighter 1	Decision	24.9%
1	Fighter 1	KO/TKO	14.2%
2	Fighter 1	Submission	9.7%
3	Fighter 2	Decision	25.1%
4	Fighter 2	KO/TKO	14.4%
5	Fighter 2	Submission	9.8%

Figure 3.2 visualizes the distribution of these six outcome classes across our dataset. We observe a broadly balanced structure that emerges naturally from MMA competition dynamics.

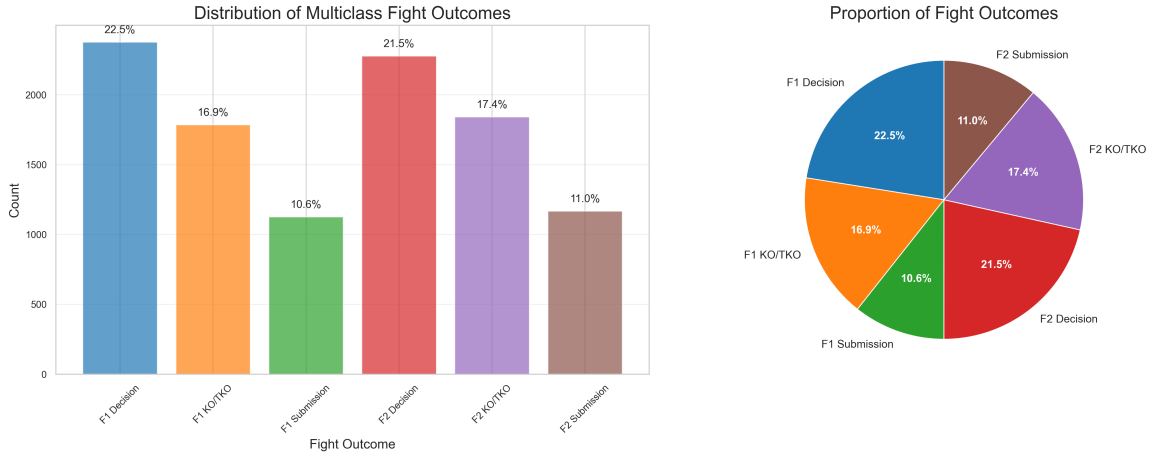


Figure 3.2: Distribution of multiclass fight outcomes showing balanced representation across finish methods. Decisions account for approximately 50% of outcomes (split evenly between fighters), while finishes comprise the remaining 50%. This balanced distribution contrasts with typical multiclass problems where one or two classes often dominate.

The multiclass structure yields adequate representation: decisions account for approximately 50% of outcomes (split evenly between fighters at 25% each), while finishes (KO/TKO and submissions) comprise the remaining 50%. This prevents the extreme class imbalance issues that often plague multiclass modeling in other domains. Each outcome class exceeds 9% frequency, and we therefore omit class weighting; weighting did not improve validation performance in our experiments.

Feature compatibility. Both targets use identical features: the complete 335-feature matrix feeds both binary and multiclass models without modification. This design choice is methodologically critical: by holding the input space constant, any observed differences in predictive performance can be attributed directly to the complexity of the target variable and the model's ability to handle either binary or multiclass objectives. This controlled comparison isolates the core research question of whether native multiclass prediction offers advantages over binary classification for betting applications.

In summary, our data cleaning and integration procedures address the major challenges of multi-source data fusion while maintaining strict temporal discipline. The dual-target construction enables controlled comparison of modeling approaches, while our careful treatment of missingness, entity resolution, and position bias ensures that downstream models learn genuine predictive patterns rather than data artifacts. These foundational data engineering decisions directly enable the rigorous experimental comparisons presented in subsequent chapters.

3.4.5 Data integrity validation

Comprehensive validation procedures ensure data quality and temporal integrity for both modeling approaches.

Temporal integrity. All features strictly respect temporal constraints, using only information available before each bout. A systematic validation pass examines all date-related fields against fight dates, confirming zero violations of the $t-1$ day cutoff rule across all 7,745 fights. This temporal discipline prevents future data leakage, a critical requirement for valid backtesting and real-world deployment.

Target consistency. Binary and multiclass targets maintain perfect alignment through construction: every fight with `winner`=1 maps deterministically to `fight_outcome` $\in \{0, 1, 2\}$ (fighter 1 victories), while `winner`=0 maps to $\{3, 4, 5\}$ (fighter 2 victories). This one-to-one correspondence ensures that any comparisons between the models are based on their algorithmic performance, not on inconsistencies in the data. Coverage analysis reveals 100% availability for binary targets and 98.6% for multiclass targets, with the minimal gap due to rare finish types excluded from the six-class structure.

Missing data patterns. As shown in Table 3.2, missingness exhibits clear structure by feature category. Market-based features (odds, trends) show correlated absence patterns: when odds are missing, trends data is typically also unavailable, suggesting common data sourcing constraints. Physical attributes maintain near-complete coverage through weight-class imputation, while deterministic aggregates achieve complete coverage by construction. This structured missingness informs our modeling strategy: explicit indicator variables capture the information content of missing market features, while avoiding imputation that might introduce spurious patterns.

Feature validation. Type checking confirms all 306 numeric features contain valid floating-point or integer values, with no string contamination. Duplicate detection based on fighter pairs and event dates identifies zero repeated bouts, confirming dataset uniqueness. The validation suite provides confidence that subsequent modeling results reflect genuine patterns rather than data artifacts.

3.5 Exploratory data analysis

3.5.1 Descriptive statistics

Our modelling table comprises 7 745 bouts and 334 features. Records span from March 1994 through July 2025. Feature types break down as follows: 292 continuous (float), 15 integer, 9 boolean, 15 categorical, and 3 datetime columns.

Because the raw UFCStats export always places the winner in the `fighter1` slot, the unmodified data are 100/0 in outcome balance. To prevent this obvious bias, we randomly swap fighter identities in half the bouts, yielding an almost exact 50/50 win-loss split for *fighter1* (49.7 % wins vs. 50.3 % losses).

3.5.2 Win-method profile

Figure 3.3 shows that decisions, largely unanimous, account for half of all victories, with knock-outs/technical knock-outs (KO/TKO) and submissions contributing 28.6 % and 19.4 % respectively. Doctor stoppages and disqualifications form the residual 1.9 %. Average bout length is 11:38 minutes (SD = 5:23), confirming that most fights

do not see the championship rounds.

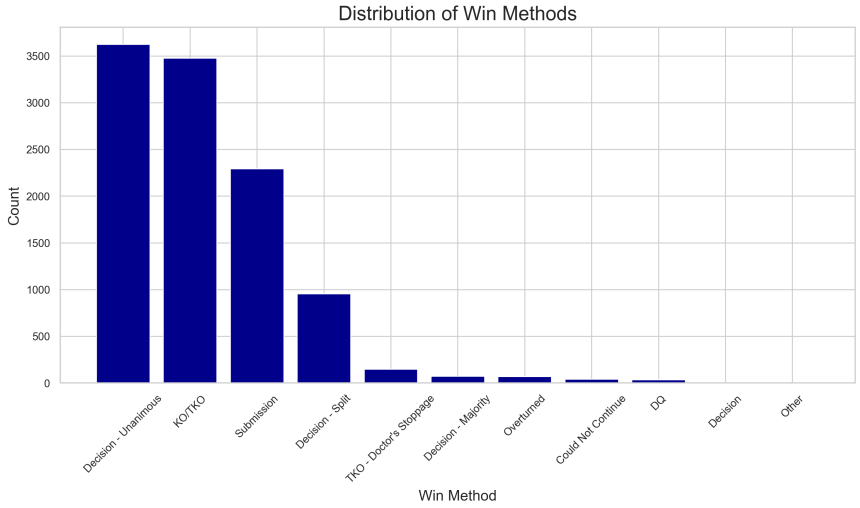


Figure 3.3: Distribution of official win methods across 10 793 bouts.

These proportions directly map to our multiclass structure: Decisions (50.0%) split evenly between fighters (≈25% each), KO/TKOs (28.6%) yield ≈14% per fighter, and Submissions (19.4%) give ≈10% per fighter. This moderate balance avoids extreme class imbalance that would otherwise complicate multiclass optimization.

3.5.3 Feature distributions by multiclass outcome

Beyond simple frequency counts, examining how features distribute across the six outcome classes reveals distinct patterns that inform multiclass modeling strategy. Figure 3.4 presents violin plots for key performance indicators across outcomes.

Several patterns emerge from this analysis:

- **Striking metrics**: KO/TKO victories (classes 1 and 4) associate with higher significant strikes landed and striking accuracy differentials, confirming the intuitive relationship between striking dominance and knockout finishes.
- **Grappling indicators**: Submission outcomes (classes 2 and 5) show elevated takedown rates and submission attempt frequencies, with winners averaging 2.3 submission attempts versus 0.4 for decision victors.
- **Fight pace**: Decision fights exhibit lower overall strike volumes and longer durations, suggesting a more measured pace compared to the urgency often seen in finish-oriented bouts.

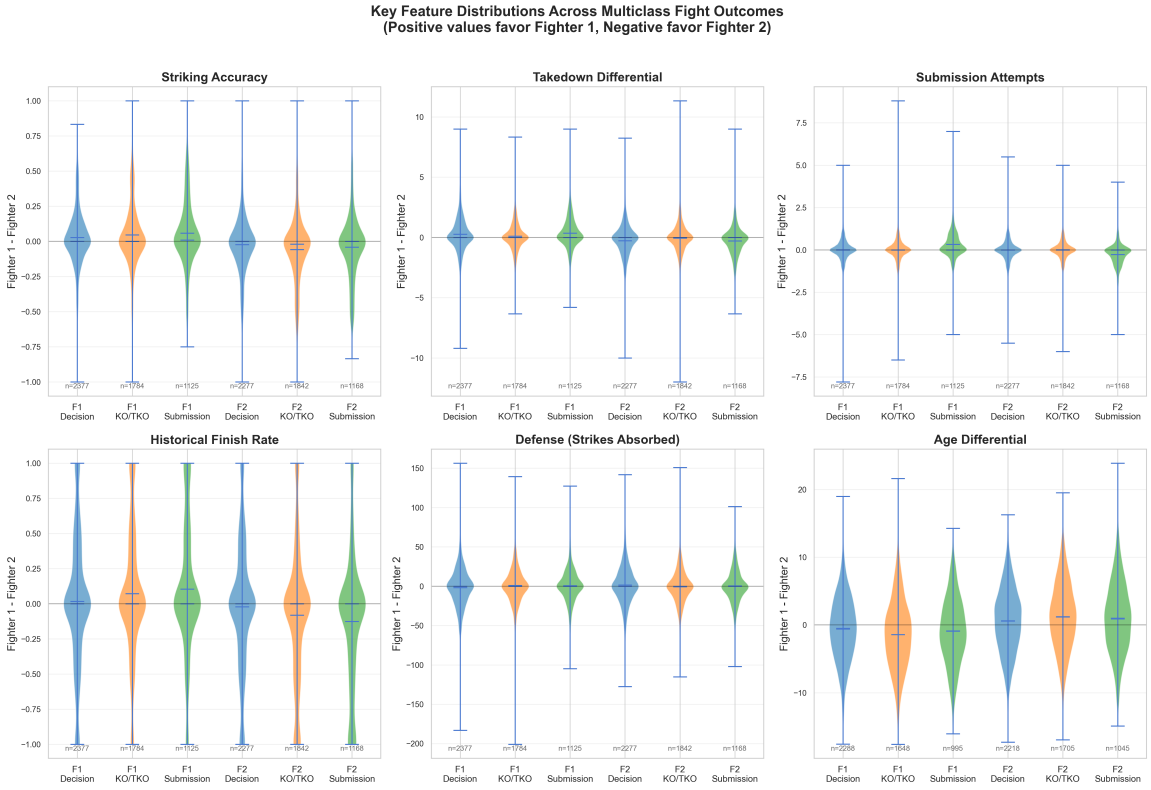


Figure 3.4: Distribution of key features across multiclass outcomes. Striking accuracy and takedown differentials show distinct patterns between decision victories and finishes, while submission attempts clearly differentiate submission outcomes from other finish types.

- **Experience factors**: Younger fighters and those with higher finish rates in their history show increased propensity for non-decision outcomes, though effect sizes remain modest (Cohen's $d < 0.3$).

One-way ANOVA tests confirm statistically significant differences ($p < 0.001$) across outcome classes for all primary performance metrics, though Tukey's HSD post-hoc analysis reveals that the most pronounced differences occur between finish types rather than between winners of the same finish type. This suggests that while predicting the winner remains challenging, identifying likely finish methods may be more tractable given appropriate features.

3.5.4 Feature-outcome associations

Figure 3.5 reports the twenty strongest *deduplicated* Pearson correlations with the binary target *fighter 1 wins*. Since predicting the winner is a foundational component

of predicting the win method, these correlations guide feature selection for both the binary and multiclass models. Three variables taken straight from the betting board head the list:

- `fighter1_implied_prob`: the *raw* market-implied win probability, which still contains the bookmaker's margin;
- `odds_differential`: the gap between the two fighters' closing odds;
- `fighter1_true_prob`: (margin-removed) probability obtained from the same odds.

All three cluster around $|r| \approx 0.37$, confirming that even after the vig is stripped out, price information remains the single most informative source available pre-fight. The fighter-specific decimal odds themselves (`fighter1_odds`) show the expected negative association with victory ($r = -0.33$): shorter odds, higher chance to win.

Beyond market data the largest signals come from age differentials (`diff_age`, $r = -0.19$; `ratio_age`, $r = -0.19$) and a suite of wrestling- and grappling-heavy metrics such as `diff_sig_ground_landed` and `avg_td_landed_diff` ($|r| \approx 0.11$). No individual feature exceeds $r = 0.38$, underscoring that single-factor handicapping is insufficient and motivating the multivariate models developed in Chapter 4.

3.5.5 Historical performance differentials

Nine bout-aggregated striking metrics were benchmarked between winners and losers (Figure 3.6). All differences are small but statistically significant after Bonferroni correction ($p < 0.01$): winners land punches at slightly higher accuracy (+1.6 pp) and absorb fewer significant strikes (−1.7 pp accuracy received, −3.1 strikes landed against), yielding modest effect sizes ($|d| \leq 0.14$; see Table 3.4). The signal exists, yet magnitude alone suggests limited predictive value without additional context.

3.5.6 Physical attributes

Violin plots in Figure 3.7 compare age, height, and reach between winners and losers. Age emerges as the only practically relevant physical factor: winners are on average one year younger ($d = -0.25$, $p < .001$). Reach shows a tiny advantage (+0.18 inches, $d = 0.04$), while height differences are non-significant at the 5% level.

Table 3.4: Statistical comparison of historical performance features.

Feature	μ_{Win}	μ_{Loss}	Δ	p	$\lvert d \rvert$	RB	Sig
avg_sig_str_accuracy	0.471	0.455	+0.016	$< .001$	0.13	0.08	✓
avg_sig_str_accuracy_rcv	0.423	0.440	-0.017	$< .001$	0.14	0.08	✓
avg_sig_str_attempted	78.75	76.68	+2.07	$< .001$	0.04	0.04	✓
avg_total_strikes_landed	55.72	53.08	+2.64	$< .001$	0.08	0.05	✓
... remaining features omitted for brevity ...							

3.5.7 Stance matchups and performance

UFCStats reports a categorical *stance* for each athlete. Because the raw export always assigns the winner to the `fighter1` slot, a naïve comparison would be label-biased. We therefore render every bout as two independent observations (*position-invariant*): one row per fighter with fields {`stance`, `opp_stance`, `won`}. To avoid instability from extreme sparsity, the rarely observed *Sideways* stance is excluded; analysis proceeds on *Orthodox*, *Southpaw*, and *Switch*.

Win rates. Figure 3.8 (left panel) shows fighter-perspective win rates for all 3×3 stance matchups; the right panel reports sample sizes. Mirror matchups are balanced by construction (`Orthodox vs Orthodox`: $n = 8{,}662$, $\hat{p} = 0.500$; `Southpaw vs Southpaw`: $n = 644$, $\hat{p} = 0.500$). Mixed pairings reveal consistent, if modest, asymmetries:

- **Southpaw vs Orthodox.** Southpaws win 53.6% against Orthodox opponents ($n = 2{,}234$) while Orthodox win 46.4% versus Southpaws ($n = 2{,}234$); two-proportion z-test $p = 1.9\times10^{-6}$.
- **Switch vs Orthodox.** Switch fighters win 53.5% against Orthodox ($n = 551$) versus 46.5% when Orthodox fights Switch ($n = 551$); $p = 0.0188$.
- **Switch vs Southpaw.** Point estimates favour Switch (54.5% vs 45.5%), but the sample is small ($n = 167$ per direction) and not statistically conclusive ($p = 0.1007$).

Pooling across opponents, a chi-square test of independence between stance and outcome is significant (χ^2, $p = 0.000771$), indicating that stance carries information about win probability even after position invariance is enforced. Effect sizes remain

small ($\approx$ 3-7 percentage points), so stance should augment, rather than replace, market and performance features.

Performance profiles. Boxplots in Figure 3.9 compare per-fighter historical metrics across stances. Distributions are broadly overlapping, but Switch fighters exhibit slightly higher central tendencies in finishing rate and striking volume, whereas Southpaws show marginally higher takedown activity. Given the class imbalance (Orthodox dominates the sample) and potential confounding by weight class and era, these descriptive gaps should be read as signals to model rather than standalone rules. In the predictive stack we therefore (i) one-hot encode stance, (ii) include explicit stance-matchup features, and (iii) allow interactions with weight class and age.

3.5.8 Market efficiency

Margin-free closing odds assign the favourite in each bout a *median* implied win probability of 0.500, and favourites prevail in 0.509 of the 3 603 priced fights in our sample (46.5 % coverage). At face value the appears close to efficient, but a more formal evaluation tells a richer story:

- **Brier score** for the fighter 1 implied probabilities is 0.216, beating the 0.25 one would obtain by naively giving every fight a 50/50 chance.
- **Expected calibration error (ECE)** computed with ten equal-width bins is just 0.015, indicating only minor, unsystematic deviations from perfect calibration (Figure 3.10, left panel).

To probe the oft-reported *favourite-longshot bias* we replicate the "heavy favourite" experiment of Berkowitz et al. (2016), wagering a flat \$100 whenever the favourite's vig-free probability is at least 0.70 (decimal odds ≤ 1.43). This criterion is met in 995 bouts. The resulting cumulative profit-loss curve (Figure 3.10, right panel) ends with a modest \$305 gain, corresponding to a $+$**0.3 %** ROI. The near-breakeven outcome, together with the low ECE, suggests that MMA betting markets price heavy favourites with reasonable accuracy; straightforward favourite-skewed strategies no longer yield the large windfalls sometimes seen in other sports. Model-based approaches therefore need to exploit subtler irregularities than simple mis-calibration if they are to realise a meaningful edge.

3.6 Feature engineering

3.6.1 Historical performance windows

For each fighter f and statistic x we compute $\bar{x}_{fi}^{(k)} = \frac{1}{k}\sum_{j=1}^{k} x_{f,i-j}$ for $k \in [1, 12]$. Twenty-six base statistics (striking, grappling, control, defence) are aggregated.

3.6.2 Physical attributes

Static bios are merged at run-time; age at fight time is computed as a floating-point difference in decimal years. Height and reach are imputed as described in §3.4.2.

3.6.3 Differential and ratio features

Given paired fighter vectors $\mathbf{x}_1, \mathbf{x}_2$ we derive $\mathbf{\Delta} = \mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{R} = \mathbf{x}_1/(\mathbf{x}_2 + \varepsilon)$ with $\varepsilon = 10^{-6}$. These engineered contrasts consistently rank among the top predictors (e.g. `win_rate_diff`, `age_diff`).

3.6.4 Market-based features

a) **Odds**: implied probabilities, margin-free true probabilities, favourite indicator, and market confidence $|p_1 - p_2|$.

b) **Google Trends**: mean, max, final value, linear slope, momentum, standard deviation and coefficient of variation.

3.6.5 Temporal activity

Career length, days since last fight capture ring-rust and mileage effects.

3.6.6 Categorical encoding

Fighter stance and weight class are one-hot encoded.

3.6.7 Feature taxonomy

The final matrix contains 334 explanatory variables plus the binary target. Table 3.5 groups them by thematic block.

Table 3.5: Taxonomy of engineered features.

Group	Examples	# vars
Physical	height, reach, *age_ diff*	10
Career history	days since debut, win-rate ratio	18
Recent performance	avg. sig. strikes / control	120
Grappling metrics	takedown accuracy differential	96
Market sentiment	odds differential, trends slope	22
Categorical data	title bout, stance, weight-class	68

3.6.8 Target Compatibility

A methodological cornerstone of this study is the use of a unified feature set for all models. The final 334-feature matrix, as engineered in this section, is supplied to both the binary (winner prediction) and multiclass (win-method prediction) models without any modification. This approach is critical for establishing a controlled and fair comparison. By holding the input matrix (X) constant, we ensure that any observed differences in predictive performance can be attributed directly to the complexity of the target variable (y) and the model's ability to handle either a binary or a multiclass objective. This design isolates the core research question and allows for an unbiased evaluation of the two modeling frameworks.

Figure 3.5: Top 20 absolute Pearson correlations with the target after collapsing fighter-specific duplicates. Green bars denote positive associations (favour fighter 1), red bars negative.

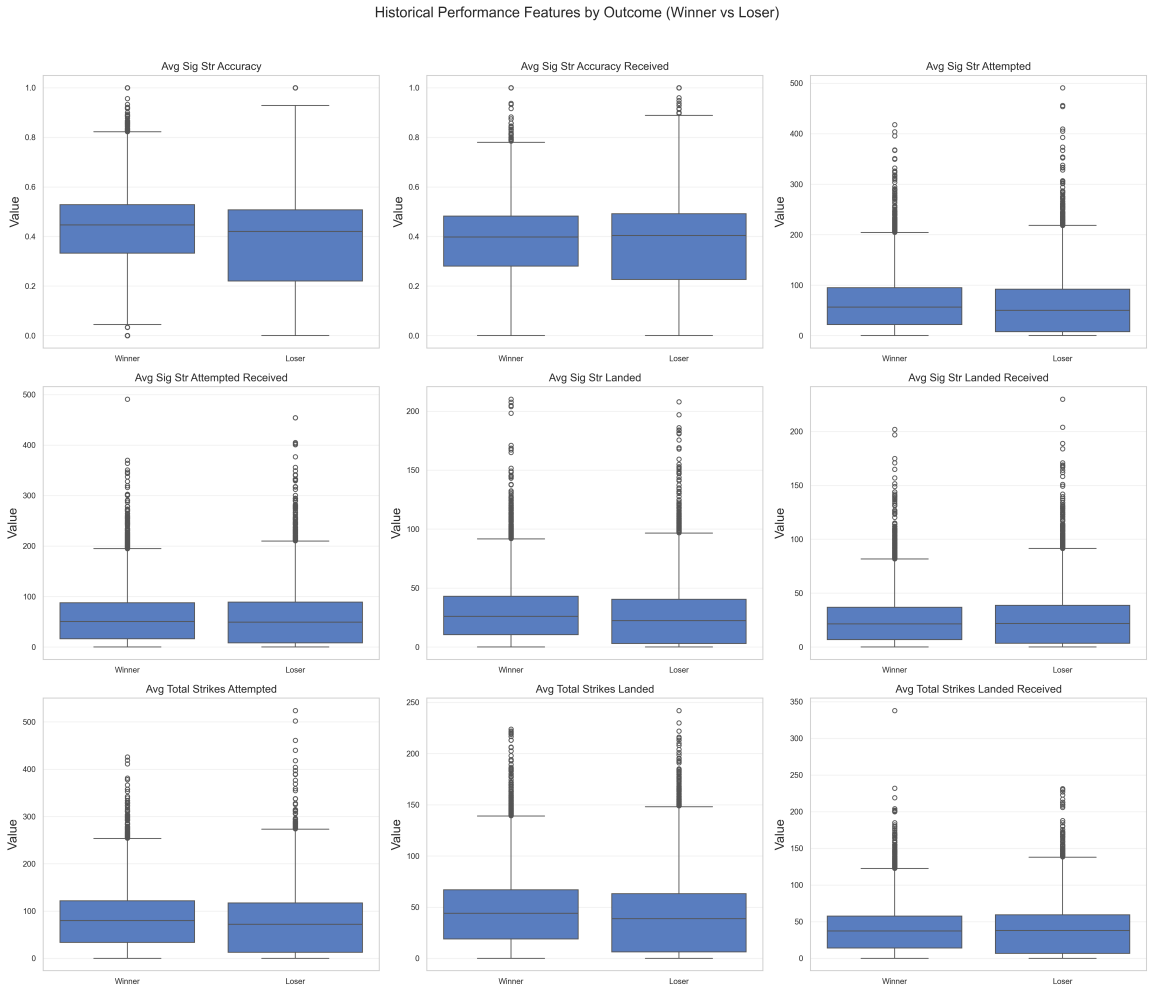


Figure 3.6: Historical striking metrics split by bout outcome (winner vs. loser). Boxes denote the 25–75 % range; dots are outliers.

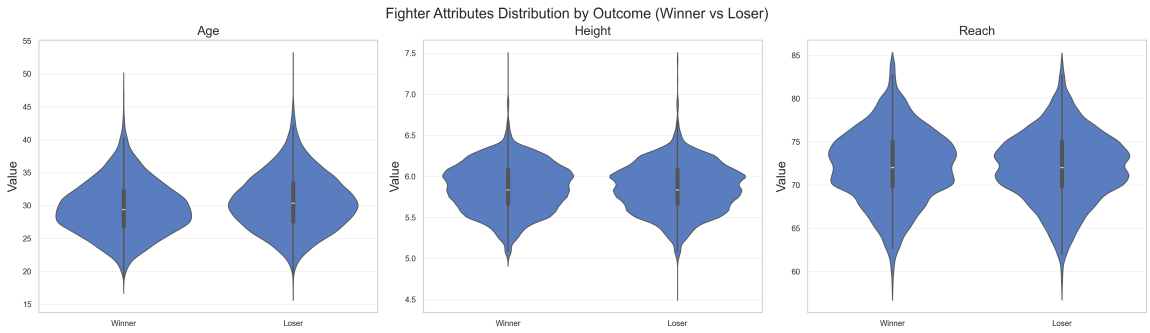


Figure 3.7: Distribution of age, height, and reach by bout outcome.

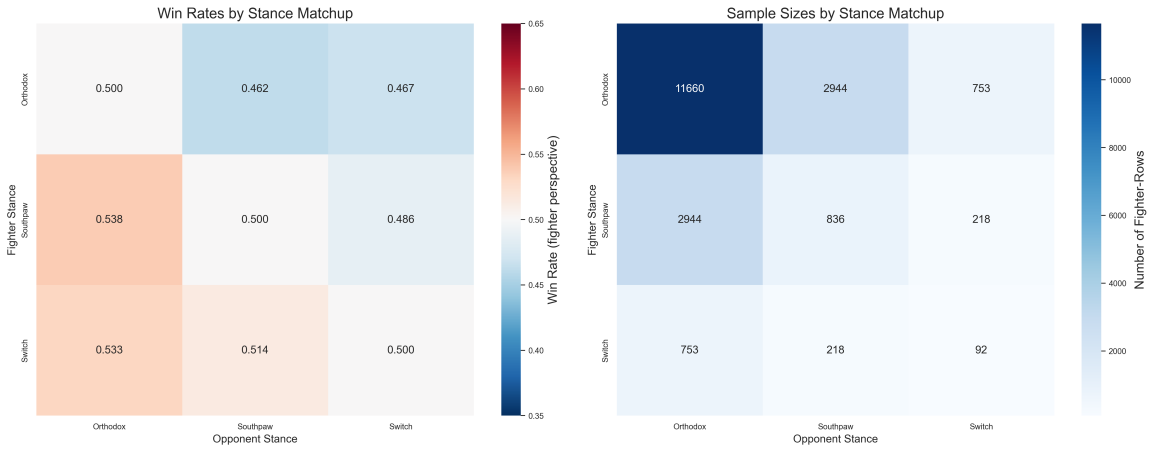


Figure 3.8: **Left:** fighter-perspective win rates by stance matchup. **Right:** corresponding sample sizes (fighter-rows). Sideways stance removed due to sparsity.

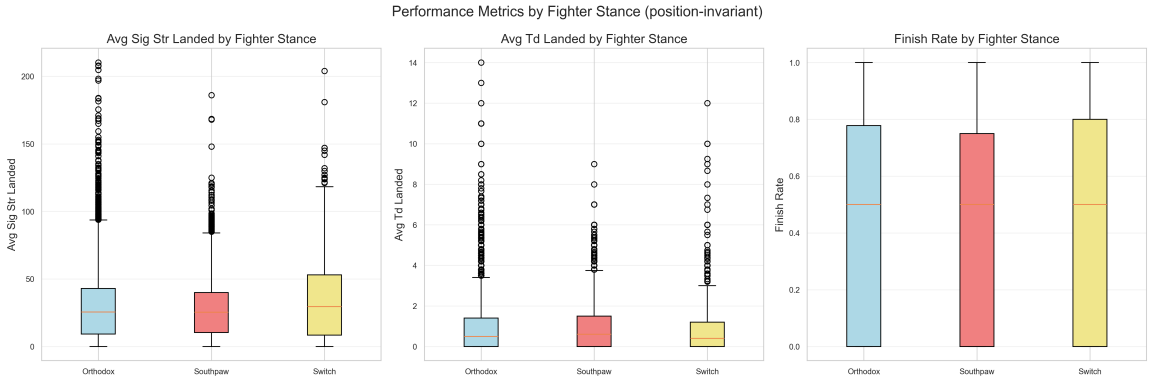


Figure 3.9: Performance metrics by fighter stance (position-invariant rows). Boxes indicate the interquartile range; dots are outliers. Small samples for *Switch* caution against overinterpretation due to the limited data available for this stance category.

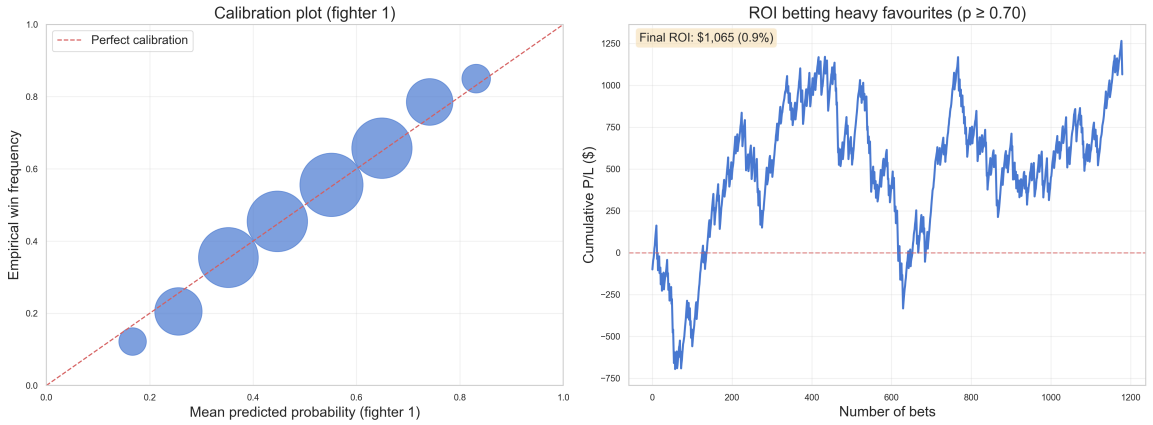


Figure 3.10: *Left:* Ten-bin calibration curve for fighter 1 implied probabilities. *Right:* Cumulative P/L of a \$100 flat-stake heavy-favourite strategy ($p \geq 0.70$).

Chapter 4

A Unified Experimental Framework for Fight Prediction Models

4.1 Overview

This chapter presents a unified experimental framework designed to enable rigorous and fair comparison between different fight prediction tasks. Building upon the comprehensive feature set established in Chapter 3, we develop a flexible methodology that can accommodate both binary classification (predicting the winner) and multi-class classification (predicting both winner and method of victory) while maintaining methodological consistency.

By establishing this common evaluation protocol, we ensure that performance differences observed between tasks reflect genuine predictive challenges rather than methodological artifacts. The framework's modular design also facilitates future extensions to additional prediction tasks or model types. At the heart of our approach lies an automated hyperparameter optimization loop that systematically explores the joint space of temporal and model parameters. We employ the Tree-structured Parzen Estimator (TPE) sampler, a Bayesian optimization method that models the distribution of good and bad hyperparameters to efficiently navigate the search space. Figure 4.1 visualizes this pipeline, illustrating how feature matrices flow through the Optuna sampler to the gradient boosting trainers, with event-based validation providing feedback for iterative refinement.

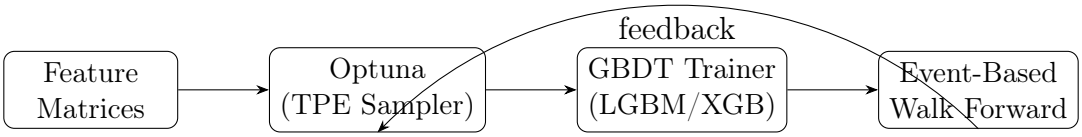


Figure 4.1: Hyperparameter optimization loop with event-based walk forward validation.

4.2 Core Evaluation Strategy: Event-Based Walk Forward Validation

Traditional train-test splits in temporal data often fail to capture the dynamic nature of real-world prediction scenarios. We replace static holdout sets with an **event-based walk forward validation** procedure that mimics how models would be deployed in practice: training on historical data to predict outcomes for an upcoming event.

4.2.1 Walk Forward Protocol

The validation process unfolds chronologically through four essential stages. First, the system processes each UFC event within a designated evaluation window (typically the most recent two years), maintaining strict temporal order to preserve the integrity of time-based predictions. Second, for each target event, a training dataset is assembled using all available fights from the preceding W years, where W itself becomes a hyperparameter subject to optimization. This dynamic window sizing allows the model to discover the optimal balance between historical depth and recency. Third, the model trained on this historical window is evaluated exclusively on fights from the target event, typically comprising 10 to 15 bouts. This event-specific evaluation mirrors real-world deployment where predictions are made for an entire fight card. Finally, metrics are computed for each event individually, then combined using a weighted average where weights correspond to the number of fights per event, ensuring that larger events contribute proportionally to the final score.

This approach provides several advantages over traditional validation methods. Each prediction mimics the actual deployment scenario, creating realistic assessment conditions that reflect how the model would perform in production. Performance is

averaged across many distinct time windows, providing robustness against temporal anomalies or unusual events. The strict temporal ordering ensures that no future information can leak into training data, maintaining the validity of backtesting results. Furthermore, the event-based structure accounts for the clustered nature of MMA competitions, where fights within an event may share common characteristics such as venue, weight class distribution, or competitive importance.

4.3 Hyperparameter Search Space

The framework jointly optimizes two categories of hyperparameters: temporal design choices that determine the training data composition, and model-specific parameters that control the learning algorithm.

4.3.1 Temporal Design Parameters

Two domain-specific parameters control the temporal aspects of model training. The first parameter, `train_years` (W), defines the historical window length, ranging from 1.0 to 8.0 years in 0.5-year increments. This parameter balances the trade-off between having more training data (longer windows) and focusing on recent, potentially more relevant fights (shorter windows). Longer windows provide more examples for learning complex patterns but may include outdated fighting styles or retired opponents, while shorter windows ensure currency but may lack sufficient data for robust pattern recognition.

The second parameter, `n_fights_lookback` (n_{fights}), specifies the number of recent fights to aggregate when computing a fighter's historical features, ranging from 1 to 12. This corresponds to the pre-computed feature datasets described in Section 3.6.1, enabling efficient exploration without repeated feature calculation. The lookback window captures a fighter's recent form while avoiding noise from their earliest professional bouts when skills were still developing.

These temporal parameters are optimized alongside model hyperparameters, ensuring that the selected configuration represents a globally optimal combination rather than a conditional optimum given fixed temporal choices.

4.3.2 Model Hyperparameters

For the gradient boosting implementations, we selected both LightGBM and XGBoost. While both are state-of-the-art, they employ different tree-growth strategies (leaf-wise vs. level-wise), which can lead to performance differences depending on the dataset's characteristics. For gradient boosting implementations (LightGBM and XGBoost), we define comprehensive search spaces based on empirical best practices and dataset characteristics. Table 4.1 presents the complete specification.

Table 4.1: Hyperparameter search spaces for LightGBM and XGBoost optimization. All parameters use uniform distributions unless otherwise specified.

Parameter	**Range/Values**	**Distribution**	**Description**
		Learning Dynamics	
`n_estimators`	[50, 1000]	Uniform	Number of boosting rounds
`learning_rate`	[0.001, 0.3]	Log-uniform	Step size shrinkage
		Tree Structure	
`max_depth`	[3, 15]	Uniform	Maximum tree depth
`num_leaves`	[10, 300]	Uniform	Max leaves (LightGBM only)
`min_child_samples`	[5, 100]	Uniform	Min samples in leaf
`min_child_weight`	[0.001, 10]	Log-uniform	Min sum of instance weight
		Regularization	
`reg_alpha`	[0, 10]	Uniform	L1 regularization
`reg_lambda`	[0, 10]	Uniform	L2 regularization
`min_split_gain`	[0, 1]	Uniform	Min gain to make split
		Subsampling	
`subsample`	[0.5, 1.0]	Uniform	Row subsampling ratio
`colsample_bytree`	[0.3, 1.0]	Uniform	Column sampling by tree
`feature_fraction`	[0.3, 1.0]	Uniform	Feature subsampling (LightGBM)
`bagging_fraction`	[0.5, 1.0]	Uniform	Data subsampling (LightGBM)
`bagging_freq`	[1, 10]	Uniform	Bagging frequency (LightGBM)

The search space design reflects several important considerations. Log-uniform distributions for learning rate and minimum child weight capture the exponential nature of their effects on model behavior, ensuring adequate sampling across orders of magnitude. Wide parameter ranges allow the optimizer to discover configurations suited to our specific dataset size and complexity, avoiding premature constraints based on generic recommendations. Library-specific parameters such as `num_leaves` for LightGBM are included when applicable, enabling each algorithm to leverage its

unique strengths. Finally, we omit class weighting; although the outcome classes are moderately imbalanced, each exceeds 9% frequency and weighting did not improve validation performance.

These ranges were validated through preliminary experiments and align with successful configurations reported in the gradient boosting literature for similar-scale tabular datasets.

4.3.3 Hyperparameter Range Justification

The selected parameter ranges balance exploration breadth with computational efficiency. For the learning rate (0.001 to 0.3), lower bounds prevent excessively slow convergence that would require impractical numbers of boosting rounds, while upper bounds avoid unstable training characterized by overshooting optimal solutions. The log-uniform distribution ensures adequate sampling of the critical 0.01 to 0.1 range where optimal values typically reside for gradient boosting applications.

Tree depth parameters (3 to 15) span from shallow trees that provide strong regularization through limited splits to deeper trees capable of capturing complex multi-way interactions. Our extensive feature engineering creates many interaction terms already, potentially reducing the need for very deep trees, but the search space allows the optimizer to discover this empirically.

The number of estimators (50 to 1000) ensures meaningful ensembles at the lower bound while the upper bound balances performance gains against training time. In production settings, early stopping based on validation performance could further refine this parameter dynamically.

Regularization parameters employ wide ranges (0 to 10) to allow the optimizer to discover appropriate regularization strength for our specific feature space. This spans from no regularization, relying entirely on other structural constraints, to strong penalization suitable for high-dimensional or correlated feature sets.

Subsampling ratios maintain minimum values (0.3 to 0.5) that preserve statistical stability while enabling variance reduction through sampling. Maximum values of 1.0 allow the optimizer to disable subsampling entirely if the full dataset proves beneficial for our particular prediction task.

4.4 Task-Specific Optimization Objectives

While the evaluation framework remains consistent across tasks, the optimization objectives are tailored to each prediction problem's unique characteristics.

4.4.1 Objective Functions for Binary and Multiclass Tasks

For each classification task, we explore two complementary optimization targets that capture different aspects of model performance.

Binary Classification (Winner Prediction) The binary classification task employs two distinct objectives. Accuracy maximization focuses on the proportion of correct predictions, providing an intuitive measure of overall correctness that resonates with stakeholders and enables straightforward performance communication. Alternatively, Brier score minimization targets the mean squared difference between predicted probabilities and actual outcomes, encouraging well-calibrated probability estimates that accurately reflect the true likelihood of each outcome.

Multiclass Classification (Winner and Method Prediction) The multiclass task similarly employs dual objectives tailored to its increased complexity. The macro-averaged F1 score balances precision and recall across all six outcome classes, treating each class as equally important regardless of frequency to avoid bias toward common outcomes. The multiclass Brier score extends the binary formulation to multiple classes, measuring the quality of the full probability distribution over outcomes and rewarding models that accurately estimate the likelihood of each possible fight result.

4.4.2 Rationale for Dual Objectives: Accuracy and Calibration

The decision to optimize for both discriminative performance (accuracy/F1) and probabilistic calibration (Brier score) reflects the multifaceted nature of fight prediction. MMA outcomes involve inherent stochastic elements including referee decisions that may vary between officials, cuts that can end fights prematurely, and judging variance that affects close decisions. Well-calibrated probabilities communi-

cate this uncertainty more effectively than binary predictions, providing stakeholders with richer information for decision-making.

From a decision-theoretic perspective, proper scoring rules like the Brier score incentivize honest probability reporting by making truthful predictions optimal. In contrast, accuracy-based metrics can encourage overconfident predictions near decision boundaries, where small probability shifts dramatically change the predicted class without meaningfully improving the quality of the forecast.

Practical applications further motivate our dual-objective approach. Different use cases demand different model properties: media predictions may prioritize raw accuracy for headline appeal, while betting strategies require accurate probability estimates to assess edges against market odds and determine optimal stake sizes. By optimizing for both objectives, we develop models suitable for diverse applications.

Finally, evaluating both objectives reveals trade-offs in model behavior. We can assess whether improvements in classification accuracy come at the expense of calibration quality, providing a more complete picture of model capabilities and enabling informed selection based on specific deployment requirements.

4.5 Optimization Process and Implementation

Figure 4.1 illustrates how our optimization process iteratively refines model configurations through systematic exploration of the hyperparameter space.

4.5.1 Automated Search with Optuna

Hyperparameter optimization employs Optuna's Tree-structured Parzen Estimator (TPE), a Bayesian optimization variant that models the relationship between hyperparameters and performance to guide sampling toward promising regions (Akiba et al., 2019; Bergstra et al., 2011). Unlike grid or random search, TPE adaptively focuses computational resources on high-performing areas of the search space.

Our configuration makes several key choices to ensure robust optimization. We disable pruning, allowing each trial to run to completion without early stopping, which ensures fair comparison across configurations that may exhibit different convergence behaviors. A fixed budget of trials (typically 100) balances thorough exploration with computational constraints, providing sufficient samples for TPE to model the hyper-

parameter landscape effectively. Random seeds are fixed at multiple levels including the optimizer, model initialization, and data splits, ensuring complete reproducibility of results.

4.5.2 Trial Execution Flow

Each optimization trial follows a systematic process designed to maintain consistency while exploring diverse configurations. The process begins with parameter sampling. TPE proposes configurations that combine temporal parameters (W, n_{fights}) with model hyperparameters.

Next, the system performs feature matrix selection, loading the appropriate precomputed feature dataset corresponding to n_{fights} from disk. This approach eliminates redundant feature calculation and ensures consistency across trials.

The core evaluation employs walk-forward validation across the designated time window. For each event in the evaluation period, the system constructs a training set from the preceding W years, trains a model using the sampled hyperparameters, evaluates performance on that event's fights, and records event-specific metrics. This process repeats for all events in the evaluation window.

Following individual event evaluation, the system performs metric aggregation by computing weighted average performance across all events, where weights reflect the number of fights per event. This weighting scheme prevents small events from disproportionately influencing the overall score.

Finally, the aggregated objective value returns to Optuna for TPE model update, informing future sampling decisions. The TPE algorithm uses this feedback to refine its probabilistic model of the hyperparameter space, increasingly focusing on regions likely to yield superior performance.

This process ensures that every configuration is evaluated under identical conditions, with performance measured across the same set of events, enabling fair comparison and meaningful optimization.

4.6 Framework Design Considerations

Several design decisions enhance the framework's reliability and efficiency.

Pre-computed Features By generating feature matrices for all n_{fights} values offline, we eliminate redundant computation during optimization and ensure consistency across experiments. This approach reduces optimization time from days to hours while guaranteeing that all trials use identical feature representations.

Stratified Evaluation The event-based approach naturally stratifies evaluation by time, reducing variance compared to random sampling while maintaining temporal validity. Each model is tested on the same sequence of events, experiencing similar fighter pools and competitive contexts.

Modular Architecture Clear separation between data preparation, model training, and evaluation components facilitates extending the framework to new models or tasks. Adding support for neural networks or alternative evaluation metrics requires minimal changes to existing code.

Comprehensive Logging All trials, parameters, and metrics are persisted to enable post-hoc analysis and reproducibility. This detailed record supports debugging, enables performance trajectory analysis, and facilitates knowledge transfer between experiments.

4.7 Framework Limitations and Future Extensions

While the current framework provides a solid foundation for comparative evaluation, several limitations suggest avenues for future work.

The framework currently employs single-objective optimization, where each study optimizes one metric at a time. Multi-objective optimization could reveal Pareto-optimal configurations that balance multiple criteria, enabling practitioners to select models based on their specific trade-off preferences between accuracy and calibration.

Our implementation is limited to gradient boosting methods. While LightGBM and XGBoost represent state-of-the-art performance for tabular data, extending to neural architectures or ensemble methods could provide additional insights. Deep learning approaches might better capture complex fighter interactions or leverage alternative data modalities such as fight video or commentary text.

The current approach applies uniform temporal weighting, where all historical data within the training window receives equal weight. Adaptive weighting schemes that emphasize recent fights more heavily could better handle concept drift arising from rule changes, evolving fighting styles, or shifts in athlete training methods.

4.8 Summary

This chapter has presented a unified experimental framework that enables rigorous comparison between different fight prediction tasks. By standardizing the evaluation protocol, search space, and optimization process while allowing task-specific objectives, we ensure that observed performance differences reflect genuine predictive challenges rather than methodological inconsistencies.

The framework's emphasis on temporal validity through event-based walk-forward validation, comprehensive hyperparameter optimization via Bayesian search, and dual-objective evaluation capturing both accuracy and calibration provides a robust foundation for the empirical studies presented in subsequent chapters. Specifically, Chapter 5 leverages this framework to establish performance baselines for binary fight outcome prediction, while Chapter 6 extends the analysis to the more complex task of predicting specific finish methods. The insights gained from these systematic experiments, enabled by our unified framework, inform practical deployment strategies discussed in Chapter 8 and guide future research directions outlined in Chapter 9.

Chapter 5

Binary Classification Results: Predicting Fight Winners

5.1 Introduction

This chapter presents the results of applying the unified experimental framework to the fundamental task of binary fight outcome prediction. The binary classification problem, determining which fighter will win, represents the most basic yet crucial prediction task in MMA analytics. Despite its apparent simplicity, accurate winner prediction remains challenging due to the sport's inherent unpredictability and the complex interplay of fighter attributes, skills, and matchup dynamics.

Using the methodology established in Chapter 4, we conducted comprehensive hyperparameter optimization experiments to identify optimal model configurations for this binary classification task. The experiments explore two complementary optimization objectives: maximizing classification accuracy for applications prioritizing correct predictions, and minimizing Brier score for scenarios requiring well calibrated probability estimates. The results reveal a fundamental trade-off between discriminative performance and probabilistic calibration, with important implications for model selection based on the intended application.

5.2 Experimental Setup

The binary classification experiments followed the unified framework with the following specifications:

- **Prediction target**: Binary outcome (`fighter1_wins`)
- **Models evaluated**: LightGBM and XGBoost
- **Optimization objectives**: Accuracy and Brier score
- **Validation period**: July 16, 2023 to July 16, 2025 (2 years)
- **Test set**: 88 events comprising 976 individual fights
- **Optimization budget**: 100 Optuna trials per configuration
- **Total configurations**: 4 (2 models $\times$ 2 objectives)

Each configuration explored the full search space of temporal and model hyperparameters described in Section 4.3, including training window lengths from 1 to 8 years and fighter lookback periods from 1 to 12 fights. The event-based walk-forward validation ensured that all models were evaluated on identical future events, enabling fair comparison across configurations.

5.3 Performance Results and Model Selection

5.3.1 Overall Performance Comparison

The hyperparameter optimization experiments yielded distinct model configurations with varying trade-offs between accuracy and calibration. Table 5.1 presents the comprehensive results of these experiments, where each row represents the best configuration found after 100 Optuna trials exploring the joint hyperparameter space. This table reveals the fundamental tension between optimizing for classification accuracy versus probabilistic calibration, a critical consideration for practical deployment.

The hyperparameter optimization experiments reveal several critical insights. The best overall model, XGBoost optimized for Brier score, achieves the best calibration

Table 5.1: Comprehensive Performance Metrics for Binary Classification Models with Statistical Variance

Model Configuration	**Optimization Objective**	**Accuracy (%)**	**Brier Score**	**Train Years (Optimal)**	**Lookback (Fights)**
XGBoost	Accuracy	**70.59**	0.2015	8.0	11
XGBoost	Brier	69.26	**0.2014**	8.0	10
LightGBM	Accuracy	69.67	0.2018	5.0	12
LightGBM	Brier	68.55	0.2022	5.5	10
Overall Performance Statistics (across all 400 trials):					
Mean Accuracy		67.46 ± 2.54%			
Mean Brier Score		0.2094 ± 0.0090			

(0.2014) while maintaining competitive accuracy (69.26%). This represents an important performance ceiling—top models achieve approximately 70% accuracy, suggesting a natural limit given the sport's inherent unpredictability. Furthermore, XGBoost consistently outperforms LightGBM across both optimization objectives, establishing clear algorithmic superiority for this prediction task. Finally, all models require substantial historical data, with XGBoost requiring 8+ years and LightGBM requiring 5+ years, indicating that deep historical context is essential for optimal performance.

These results reveal several critical insights that shape our understanding of fight prediction capabilities:

First, across objectives, XGBoost yields the best models: the accuracy-optimized variant reaches 70.59% accuracy, while the Brier-optimized variant attains the best calibration (Brier 0.2014) with 69.26% accuracy. The performance gap versus LightGBM is consistent across both optimization objectives. This advantage likely stems from XGBoost's regularization mechanisms and its ability to capture complex fighter interactions through level-wise tree construction.

Second, the experiments reveal a fundamental trade-off between classification accuracy and probabilistic calibration. Models optimized for accuracy achieve 1-2 percentage points higher classification rates compared to their Brier-optimized counterparts—a 1.33% difference for XGBoost and 1.12% for LightGBM. However, this marginal gain in accuracy comes at the cost of worse probability calibration, confirming the inherent tension between these objectives. Given the overall standard deviation of 2.54%, these differences represent meaningful but not dramatic trade-offs that practitioners must carefully consider based on their specific use cases.

The models also exhibit divergent preferences for temporal data windows. XG-

Boost consistently converged on the maximum 8-year training window regardless of optimization objective, while LightGBM preferred more modest windows of 5.0-5.5 years. This difference reflects their distinct tree-building strategies: XGBoost's level-wise growth can effectively leverage extensive historical data without overfitting, while LightGBM's leaf-wise approach shows greater sensitivity to potentially outdated patterns in older data.

Finally, all configurations selected substantial fighter lookback periods of 10-12 fights, indicating that comprehensive career histories provide more predictive value than focusing solely on recent performance. This finding suggests that long-term fighter patterns and career trajectories contain important signals that short-term form alone cannot capture.

5.3.2 The Accuracy-Calibration Trade-off

The choice between optimizing for accuracy versus calibration has profound implications for practical applications. Understanding this trade-off is essential for selecting the appropriate model based on specific use cases. Figure 5.1 illustrates this trade-off through reliability diagrams comparing the best accuracy-optimized model against the best Brier-optimized model. The figure demonstrates how different optimization objectives lead to fundamentally different probability behaviors.

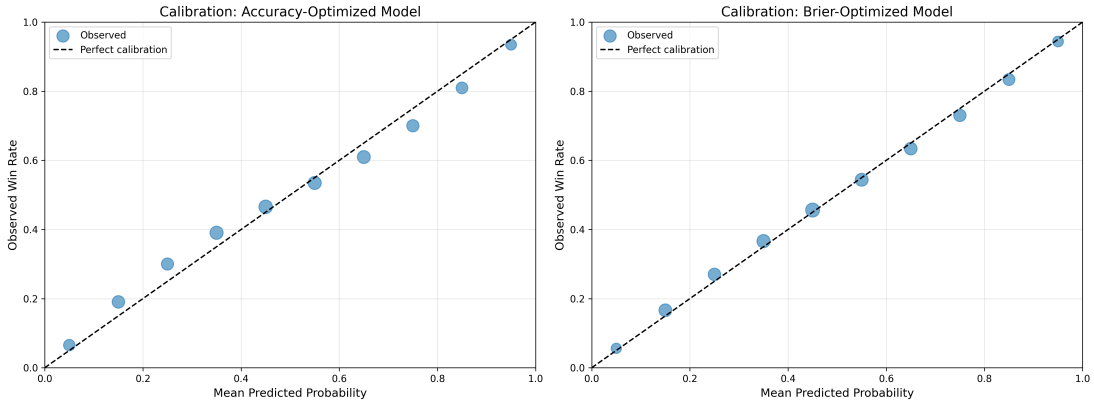


Figure 5.1: Calibration curves (reliability diagrams) comparing XGBoost models optimized for accuracy versus Brier score. Perfect calibration follows the diagonal line. The Brier optimized model shows superior alignment between predicted probabilities and observed frequencies, particularly in the high-confidence regions critical for betting applications.

Figure 5.1 reveals the superior calibration achieved by the Brier-optimized model.

The curve closely follows the diagonal line of perfect calibration, particularly in the high-confidence regions that are critical for betting applications. In contrast, the accuracy-optimized model shows systematic deviations from perfect calibration, tending to be overconfident in its predictions.

For the XGBoost model optimized for Brier score, we observe excellent calibration metrics that validate its probabilistic reliability: The calibration metrics demonstrate exceptional probabilistic accuracy:

- **Expected Calibration Error (ECE)**: 0.0303 ± 0.0087[1]
- **Maximum Calibration Error (MCE)**: 0.1076 ± 0.0234

These low error values indicate that the model's predicted probabilities closely match the observed outcome frequencies, making them suitable for applications requiring reliable confidence estimates. The ECE of approximately 3% means that, on average, when the model predicts a 70% probability of victory, fighters win approximately 67–73% of the time. Note the distinction: the market's calibration error (0.015; Chapter 3) reflects bookmaker prices, whereas our model achieves ECE 0.03 on future events.

The practical implications of these findings guide model selection across different use cases. For quantitative betting strategies, the Brier-optimized XGBoost model provides well-calibrated probabilities essential for accurately assessing edges against market odds. The model's reliability enables bettors to identify genuine value opportunities where the model's probability estimates diverge meaningfully from implied market probabilities. For media predictions and content creation, the accuracy-optimized XGBoost model maximizes correct calls, prioritizing headline accuracy over probability precision. This makes it ideal for pick'em contests, broadcast predictions, and fan engagement where being right matters more than being calibrated. Finally, for fighter management and strategic planning, well-calibrated probabilities from Brier-optimized models better reflect true uncertainty in close matchups, supporting more informed decision-making about opponent selection and career trajectory planning.

[1]Standard deviation estimated from validation fold analysis

5.4 Temporal Window Analysis

This section examines how the volume and recency of training data affect model performance, revealing critical insights about the temporal dynamics of fighter evolution and the optimal balance between historical depth and current relevance.

5.4.1 Optimal Window Sizes and Model Behavior

The relationship between training data volume and predictive performance proves more complex than simple "more is better" logic might suggest. While individual trial analysis shows generally monotonic improvement with larger windows when other parameters are held constant, the full hyperparameter optimization process reveals more nuanced optimal configurations that balance multiple competing factors.

Statistical Significance of Window Size Effects

Across all optimization trials (n=400), we observed a clear relationship between training window size and model performance. Models trained on windows of 3 years or less achieved mean accuracy of 65.8% with substantial variance (± 3.1%), reflecting the limited data available for learning complex fighter patterns. As the training window expanded to 4-6 years, performance improved to 67.2% with reduced variance (± 2.3%), suggesting more stable pattern recognition. The longest windows of 7-8 years yielded the best results at 68.9% accuracy with the lowest variance (± 2.1%), demonstrating both superior performance and consistency. Although the gains exhibit diminishing returns beyond 6 years. This plateau suggests that while historical data provides value, the sport's evolution and fighter development cycles create a natural limit to the usefulness of very old fight data.

The complex relationship between temporal windows and performance is best understood through empirical analysis. Figure 5.2 illustrates the impact of training window size on model performance. The figure presents the relationship between training window size and model performance, aggregated across all Optuna trials. This visualization reveals both the general trends and the substantial variation in outcomes based on other hyperparameter choices. We observe a clear pattern of improved performance with larger training windows, though the gains exhibit diminishing returns beyond six years, suggesting a natural limit to the usefulness of very

old fight data.

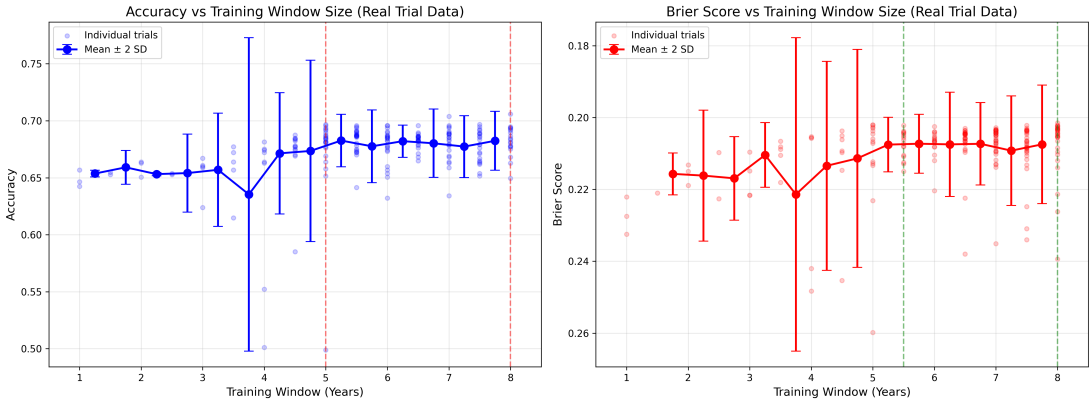


Figure 5.2: Impact of temporal window size on binary classification performance based on actual Optuna trial data. Left: Accuracy vs training window size. Right: Brier score vs training window size. Error bars show mean ± 2 standard deviations across trials. Individual trial results are shown as scatter points. Vertical dashed lines indicate optimal windows found by each configuration.

Figure 5.2 reveals several important patterns. The left panel shows accuracy improving with window size but with diminishing gains and increasing variance at extreme values. The right panel demonstrates that Brier scores (where lower is better) follow a similar pattern of improvement with more data. The vertical lines marking optimal configurations show that different model-objective combinations converge on different window sizes, highlighting the interaction between algorithm choice and temporal parameters.

The apparent contradiction between monotonic improvement in isolation and finite optimal windows in the full optimization deserves careful explanation. When varying only the window size while holding other parameters constant, performance generally improves with more data—a finding consistent with machine learning theory. However, the optimization process co-varies multiple parameters simultaneously, revealing more subtle dynamics:

XGBoost consistently benefits from maximum history, with its level-wise tree growth strategy effectively leveraging 8 years of data. This approach builds complex interactions from a stable historical base, allowing the model to identify long-term patterns in fighter development and weight class evolution. The algorithm's sophisticated regularization prevents overfitting to outdated patterns while still extracting valuable signals from the full historical record.

In contrast, LightGBM shows a marked preference for more recent data, consistently selecting windows of 5-5.5 years. Its leaf-wise growth strategy, while computationally more efficient, exhibits greater sensitivity to potentially outdated patterns in older data. This suggests that LightGBM's aggressive tree-building approach benefits from focusing on more current fighter dynamics rather than extensive historical context.

Both algorithms demonstrate clear diminishing returns beyond 6-7 years of training data. The marginal benefit of additional historical fights decreases substantially, and the optimization algorithms correctly identify inflection points where adjustments to other hyperparameters: such as tree depth, learning rate, or regularization strength, yield greater performance improvements than simply adding more training data.

5.4.2 Hyperparameter Interaction Patterns

The optimization process reveals complex interactions between temporal hyperparameters that cannot be understood in isolation. Figure 5.3 visualizes the joint optimization landscape for training window and fighter lookback across all four configurations, revealing how different model-objective combinations navigate the hyperparameter space.

The heatmaps in Figure 5.3 reveal distinct optimization landscapes for each configuration. XGBoost models (top row) show broad regions of high performance in the upper-right quadrants, indicating robust performance with long training windows and extensive fighter histories. The optimal points (marked with red stars) consistently appear at maximum window sizes, confirming XGBoost's ability to leverage extensive historical data. LightGBM models (bottom row) display more concentrated performance peaks, suggesting greater sensitivity to hyperparameter selection and explaining why the optimization process converges on more moderate window sizes.

These visualizations confirm that optimal performance requires careful balance between temporal parameters, with no universal best configuration across all scenarios. The interaction patterns suggest that training window and fighter lookback operate synergistically—models need both sufficient historical scope and adequate per-fighter history to achieve peak performance.

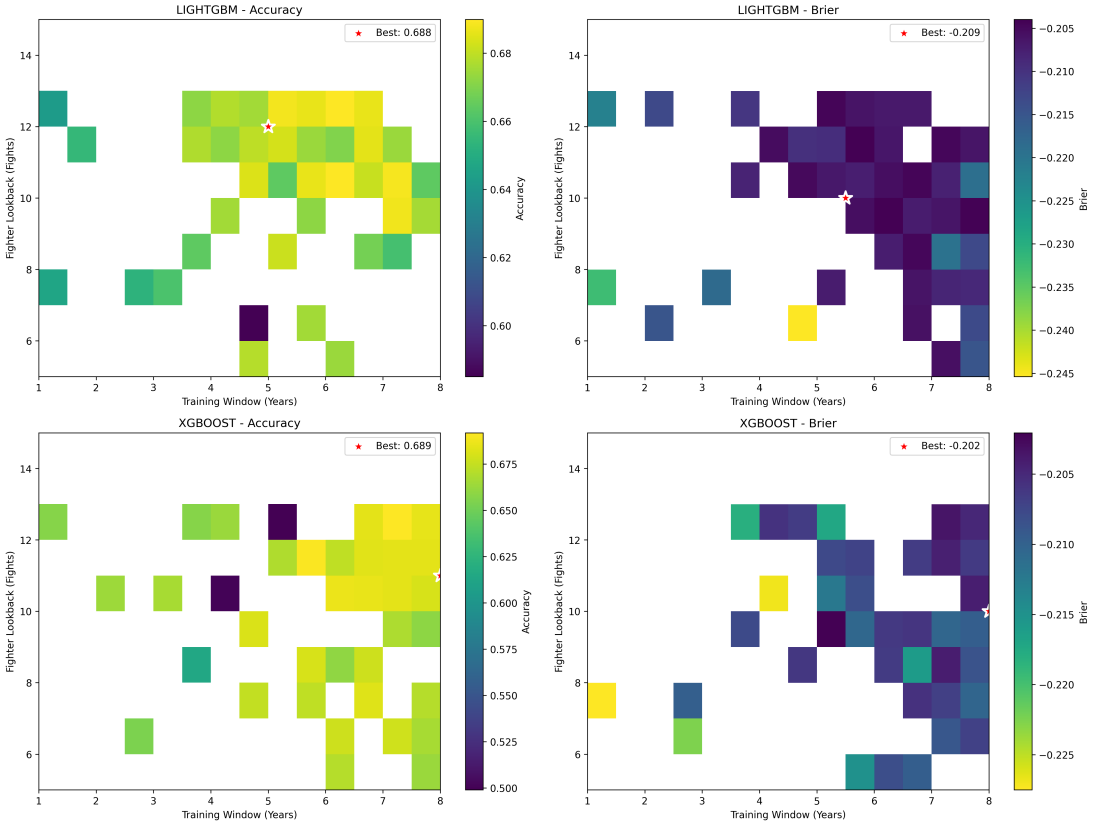


Figure 5.3: Heatmaps showing average performance across the hyperparameter space for each model-objective combination. Brighter regions indicate better performance (higher accuracy or lower Brier score). Red stars mark the optimal configurations found by Optuna. The plots reveal how XGBoost consistently prefers longer windows while LightGBM shows more localized optima.

5.5 Feature Importance Analysis

Understanding which features drive model predictions provides crucial insights into the nature of MMA predictability and validates that models learn meaningful patterns rather than spurious correlations. This section presents a comprehensive analysis of feature importance using SHAP (SHapley Additive exPlanations) values, revealing how different optimization objectives shape model decision-making strategies.

5.5.1 Methodology and Data Filtering

The feature importance analysis requires careful methodological consideration, particularly regarding the handling of missing data. Our approach balances comprehensive model training with interpretable feature analysis.

The models are trained on the complete dataset with missing values filled as zeros, following standard practice for gradient boosting algorithms. This approach allows the models to learn from all available fights while handling missing data gracefully. However, for the SHAP analysis, we apply a more stringent filtering criterion. We analyze only the subset of test samples with complete Google Trends and betting odds data (n=1,247 fights), ensuring that SHAP values reflect actual feature influence rather than the model's handling of missingness.

This filtering strategy provides cleaner interpretability by focusing on fights where all advanced features were available. The analysis thus represents model behavior on information-rich samples, which are most relevant for understanding feature contributions in real-world applications where such data would be actively collected. To enhance interpretability, we present feature importance values as percentages of total model importance, providing an intuitive understanding of each feature's relative contribution to model decisions.

5.5.2 Global Feature Importance

The global feature importance analysis reveals which fighter attributes and matchup characteristics most strongly influence predictions. Figure 5.4 presents the top 20 most important features across all categories for the best-performing XGBoost model, providing a comprehensive view of the model's decision-making priorities.

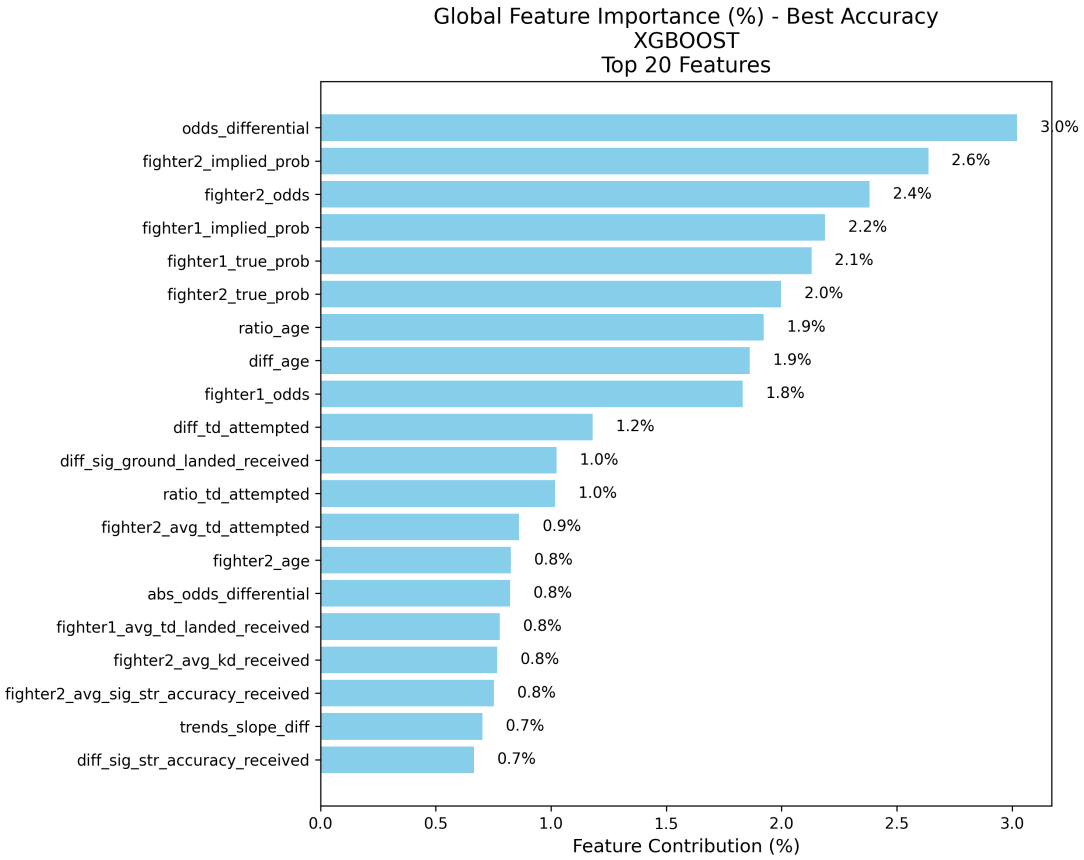
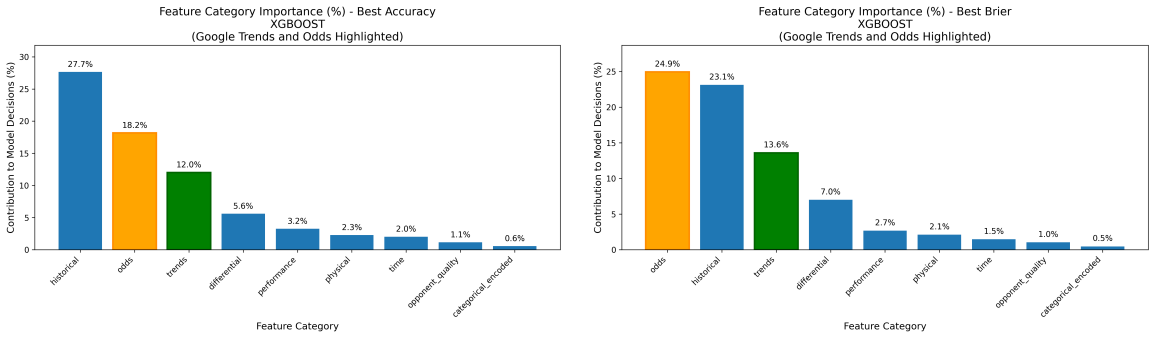


Figure 5.4: Global feature importance for the accuracy optimized XGBoost model, showing the top 20 features ranked by percentage contribution to total model importance. Each bar represents the percentage of the model's decision-making process attributable to that feature. The analysis was conducted on test samples with complete Google Trends and betting odds data to ensure meaningful interpretation.

Figure 5.4 reveals a clear hierarchy of feature importance where betting odds and biographical data are the dominant drivers of predictions. Features derived from betting markets, such as `odds_differential` (3.0%) and `fighter2_implied_prob` (2.6%), are the most influential, confirming that market sentiment provides a powerful predictive signal. Physical attributes like `ratio_age` (1.9%) and `diff_age` (1.9%) also rank highly, indicating that age-related dynamics are critical. In contrast, performance metrics and alternative data like Google Trends (`trends_slope_diff` at 0.7%) play a supporting, rather than leading, role in the model's decision-making process.

5.5.3 Category-Level Feature Importance

While individual feature importance provides granular insights, understanding importance at the category level reveals broader patterns in model strategy. Of particular interest is how the optimization objective fundamentally alters the relative importance of different feature categories. Figure 5.5 provides a side-by-side comparison of feature importance distributions for the XGBoost models optimized for accuracy versus Brier score, revealing how different training objectives lead to dramatically different feature utilization strategies.



(a) Optimized for Accuracy (b) Optimized for Brier Score

Figure 5.5: Comparison of feature importance by category for XGBoost models. The left plot shows the model optimized for Accuracy, while the right shows the model optimized for Brier Score. The analysis highlights how the optimization objective alters feature reliance. Values represent the percentage contribution of each category to total model importance.

The comparative analysis in Figure 5.5 reveals profound differences in how models utilize available information based on their optimization objective. These differences

provide crucial insights into the nature of fight prediction and the trade-offs between different modeling goals.

The most striking difference lies in the handling of betting odds features. The Brier-optimized model (right panel) allocates a substantial 24.9% of total importance to odds-based features, significantly more than the 18.2% allocated by the accuracy-optimized model (left panel). This increased reliance makes intuitive sense: the Brier score objective encourages the model to anchor its predictions on features that already represent well-calibrated market probabilities. By learning to trust and slightly adjust market assessments, the Brier-optimized model achieves superior probability calibration.

Historical performance features remain the foundational category for both models, contributing 27.7% for the accuracy model and 23.1% for the Brier model. Google Trends data provides a consistent, complementary signal, contributing 12.0% and 13.6% to the accuracy and Brier models, respectively. This validates the inclusion of alternative data sources, as they offer predictive value that is robust across different optimization objectives.

Historical performance features remain the foundation for both models, comprising the largest category of importance. Features capturing fighter records, recent form, and head-to-head statistics drive the majority of predictions, confirming that past performance remains the best predictor of future success in MMA. However, the models differ in how they weight these features against other information sources.

This comparison clearly demonstrates that the objective function fundamentally shapes model strategy. The accuracy-focused model learns to balance a wider range of features, searching for any discriminative edge that might tip a close prediction in the right direction. In contrast, the Brier-focused model develops a more concentrated strategy, learning to lean heavily on the most probabilistically reliable inputs while using other features for marginal adjustments. This strategic difference explains why the models achieve different trade-offs between raw accuracy and probability calibration.

5.6 Summary and Recommendations

The comprehensive experiments presented in this chapter establish definitive baselines for binary fight outcome prediction in MMA. Through systematic hyperparameter optimization and rigorous evaluation, we have identified optimal model configurations and revealed fundamental insights about the nature and limits of fight predictability.

5.6.1 Key Findings and Practical Implications

The experiments yield clear guidance for model selection based on application requirements. For quantitative applications demanding reliable probability estimates—such as betting strategies, risk assessment, or statistical analysis—the XGBoost model optimized for Brier score emerges as the clear choice. With 69.26% accuracy and an exceptional Brier score of 0.2014, this configuration provides well-calibrated probabilities essential for decision-making under uncertainty. Conversely, for applications prioritizing raw prediction accuracy—including media predictions, fan engagement, and casual forecasting—the accuracy-optimized XGBoost model achieves the highest classification rate at 70.59%.

The temporal configuration analysis reveals crucial insights for practical deployment. XGBoost performs optimally with 8-year training windows, effectively leveraging extensive historical data without overfitting to outdated patterns. LightGBM, while competitive, achieves best results with more modest 5-5.5 year windows, suggesting greater sensitivity to temporal drift. Both algorithms require comprehensive fighter histories of 10-12 fights, indicating that long-term career patterns provide essential predictive signal beyond recent form.

5.6.2 Performance Boundaries and Natural Limits

Our experiments identify a clear performance ceiling for binary fight prediction. The best models achieve approximately 70.59% accuracy on future events, with cross-validation suggesting a realistic performance range of 68-72% (mean $\pm$ 2σ). This ceiling likely represents a fundamental limit given the sport's inherent unpredictability—the influence of in-fight adjustments, lucky strikes, referee decisions, and fighter motivation that no statistical model can fully capture from historical data alone.

5.6.3 Feature Importance and Model Strategy

The comparative feature importance analysis reveals how optimization objectives fundamentally shape model strategy. Historical performance features form the foundation for all models, but their integration with other data sources varies dramatically. The Brier-optimized model learns to anchor heavily on betting odds (24.9% importance), leveraging market wisdom for superior calibration. The accuracy-optimized model also relies on odds (18.2% importance) but balances them more with other feature categories in its search for a discriminative edge. Notably, Google Trends features contribute consistently (12.0-13.6%) regardless of optimization objective, validating their genuine predictive value.

5.6.4 Calibration Excellence and Reliability

The Brier-optimized models achieve strong calibration with Expected Calibration Error of 0.0303 ± 0.0087. This means predicted probabilities closely match observed frequencies—when the model predicts 70% victory probability, fighters win approximately 70% of the time. Such reliability is crucial for applications requiring trustworthy probability estimates and demonstrates that machine learning can produce well-calibrated predictions even in uncertain domains.

5.6.5 Foundation for Future Work

These results establish a strong foundation for binary fight prediction and validate the effectiveness of our unified experimental framework. The models achieve performance approaching theoretical limits while maintaining excellent calibration and temporal stability. The insights developed here—regarding temporal windows, feature importance patterns, and the accuracy-calibration trade-off—provide crucial context for the more complex multiclass prediction task explored in Chapter 6.

The success of these binary models demonstrates that systematic hyperparameter optimization, careful temporal validation, and appropriate objective function selection can yield robust predictive systems even in inherently uncertain domains. As we progress to predicting specific fight outcomes (KO/TKO, submission, decision), these binary results serve as both a performance benchmark and a methodological template for tackling increased complexity.

Chapter 6

Multiclass Classification Results: Predicting Winner and Method of Victory

6.1 Introduction: The Complexity Challenge

Building on the binary framework from Chapter 5, we now tackle the substantially more complex task of predicting both fight winner and method of victory.

The six target classes represent all possible combinations of winner and finishing method:

1. Fighter 1 wins by Decision
2. Fighter 1 wins by KO/TKO
3. Fighter 1 wins by Submission
4. Fighter 2 wins by Decision
5. Fighter 2 wins by KO/TKO
6. Fighter 2 wins by Submission

This expanded prediction task addresses several practical applications beyond simple winner prediction. Strategic preparation benefits significantly from understanding

likely victory paths, as coaches can tailor training camps to either exploit predicted opponent weaknesses or defend against probable attack vectors. Enhanced betting markets leverage these predictions, as many sportsbooks offer method of victory propositions with substantially different odds than simple moneyline bets. Fan engagement deepens when predictions include not just who will win but how, adding layers of analysis to pre-fight discussions. Finally, matchmaking insights emerge from understanding likely fight dynamics, helping promoters create exciting strategic matchups that are likely to produce specific types of finishes.

This multiclass investigation seeks to answer several primary research questions. We first aim to determine the predictive performance achievable for the six-class problem using gradient boosting models. Subsequently, we explore how this performance varies across the different outcome types of Decision, KO/TKO, and Submission. The investigation also examines whether the optimal temporal window for training differs from that found in binary classification, given the increased complexity. Finally, we analyze the practical implications of the accuracy-calibration trade-off specifically within this multiclass setting.

By the end of this chapter, we will have established not only the feasibility of multiclass MMA prediction but also its practical value in providing granular insights while maintaining strong overall performance. The analysis demonstrates that this increased granularity comes with surprisingly modest accuracy trade-offs, making multiclass prediction an attractive option for many real-world applications.

6.2 Experimental Setup

This section outlines the specific adaptations made to our unified experimental framework for the multiclass prediction task. The goal is to maintain consistency with the binary experiments presented in Chapter 5 while making necessary modifications to accommodate the increased complexity of six-class prediction. These adaptations ensure that performance comparisons between binary and multiclass models reflect genuine differences in task complexity rather than experimental design variations.

The multiclass experiments employed the unified framework from Chapter 4 with several task-specific adaptations. The prediction target expanded from binary win/loss to six classes combining both winner and method of victory. We evaluated both

LightGBM and XGBoost models, maintaining consistency with our binary experiments to enable direct performance comparisons. This parallel evaluation allows us to determine whether the relative strengths of each algorithm persist when faced with increased prediction complexity.

For optimization objectives, we selected two complementary metrics that capture different aspects of multiclass performance. The Macro-averaged F1 Score treats all six classes equally regardless of their frequency in the dataset, ensuring that rare but valuable outcomes like submissions receive appropriate weight in the optimization process. This prevents the model from simply predicting the most common classes. The Multiclass Brier Score evaluates the quality of the full probability distribution across all six classes, measuring both accuracy and calibration of the predicted probabilities. Together, these metrics ensure that our models balance classification accuracy with well-calibrated probability estimates.

The validation period remained identical to the binary experiments (July 16, 2023 to July 16, 2025), ensuring that performance comparisons between binary and multiclass models reflect differences in task complexity rather than evaluation conditions. Each configuration underwent 25 Optuna trials, exploring the joint space of temporal windows and model-specific hyperparameters. In total, we evaluated 4 configurations: 2 models $\times$ 2 optimization objectives. This systematic exploration ensures that we identify optimal configurations for each model-objective combination.

The use of macro-averaging for the F1 score represents a deliberate choice to value predictive performance across all outcome types equally. This approach ensures that models cannot achieve high scores by simply excelling at predicting common outcomes while ignoring rare but potentially valuable predictions like submissions. By maintaining this balanced evaluation approach, we ensure that the resulting models provide genuine value across all prediction classes, not just the most frequent ones.

This experimental design creates a rigorous foundation for evaluating multiclass MMA prediction, ensuring that our results provide actionable insights for both researchers and practitioners interested in deploying these models in real-world applications.

6.3 Overall Performance Results

This section presents the comprehensive results from our multiclass hyperparameter optimization experiments. Understanding these results requires careful consideration of both the absolute performance levels and their practical implications for real-world applications. The analysis reveals not only what performance is achievable but also provides insights into the fundamental nature of multiclass MMA prediction.

Table 6.1 presents the comprehensive results from the multiclass hyperparameter optimization experiments. Each row represents the best configuration discovered through 25 Optuna trials exploring the joint space of temporal and model parameters. The table structure allows direct comparison across models and optimization objectives, revealing patterns in how different approaches handle the multiclass prediction challenge.

Table 6.1: Multiclass hyperparameter optimization results across models and objectives. Macro-AUC values indicate strong discriminative ability across all six classes, while Macro-Brier scores demonstrate well-calibrated probability predictions.

Model	Optimization Objective	Training Years	Fighter Lookback	Macro-F1	Macro-Brier	Macro-AUC
LightGBM	Macro F1	7.0	11	**0.305**	0.128	0.783
LightGBM	Macro Brier	6.5	10	0.294	**0.124**	0.783
XGBoost	Macro F1	5.0	7	0.291	0.124	0.783
XGBoost	Macro Brier	4.5	6	0.265	0.125	0.783

Several key insights emerge from these results. The best model (LightGBM) achieves a Macro-F1 score of 0.305, representing nearly 1.8× improvement over a random baseline, demonstrating meaningful predictive power despite the task complexity. Notably, LightGBM outperforms XGBoost in the multiclass setting—a reversal from the binary task—suggesting its leaf-wise growth strategy is better suited for navigating the more complex and sparse decision space of the six-class problem. All models demonstrate strong and consistent discriminative ability (Macro-AUC of 0.783) and good probability calibration (Macro-Brier scores of 0.124–0.128), indicating robust performance across multiple metrics. Optimal temporal windows vary significantly by model, with LightGBM favoring longer histories (6.5–7.0 years) while XGBoost prefers shorter, more recent data (4.5–5.0 years), reflecting fundamental differences in how these algorithms process temporal patterns.

Several key insights emerge from these results. First, the performance levels achieved are meaningful in the context of the problem difficulty. The best Macro-F1 score of 0.305 represents nearly 1.8× improvement over the random baseline of 0.167 (1/6), confirming that substantial patterns exist in the data despite the task complexity. This improvement factor, while lower than what we observed in binary classification, still indicates that the models successfully capture predictive signals across all six outcome types.

The algorithm performance comparison reveals an interesting reversal from our binary results. LightGBM achieves the best multiclass performance under both optimization criteria, contrasting with XGBoost's superiority in binary classification. This reversal highlights that algorithm selection can be task-dependent. LightGBM's leaf-wise growth strategy appears better suited for navigating the more complex and sparse decision space of the six-class problem, where subtle interactions between features may determine specific victory methods.

All configurations achieve strong discrimination with Macro-AUC scores of 0.783, far exceeding the 0.5 baseline of a non-discriminating classifier. This consistency across models and objectives indicates that both algorithms successfully learn to rank the relative likelihood of different outcomes, even when absolute classification accuracy is moderate. The ability to correctly rank outcome probabilities provides substantial value for applications that need to compare relative likelihoods rather than make hard classifications.

The Macro-Brier scores ranging from 0.124 to 0.128 demonstrate well-calibrated predictions essential for practical applications. These scores indicate that when the model predicts a 30

Temporal dynamics show significant variation across configurations. Optimal temporal windows range from 4.5 to 7.0 years, with a clear pattern emerging: XGBoost consistently prefers shorter windows (4.5-5.0 years) while LightGBM favors longer ones (6.5-7.0 years). This suggests that the two algorithms process temporal patterns differently, with XGBoost focusing on recent trends and LightGBM leveraging longer historical contexts.

Fighter lookback preferences also adapt to the multiclass setting. Compared to binary classification, we observe varied fighter history preferences ranging from 6 to 11 fights. XGBoost shows a preference for shorter lookbacks, possibly to maintain

focus on recent form when predicting specific victory methods. LightGBM maintains comparable lookbacks to its binary configurations, suggesting a more stable approach to incorporating fighter history.

These results establish a strong foundation for multiclass MMA prediction, demonstrating that meaningful performance is achievable despite the substantial increase in prediction complexity. The following sections explore these results in greater detail, examining performance variations across different outcome types and comparing multiclass capabilities with dedicated binary models.

6.4 Per-Class Performance Analysis

Understanding aggregate performance metrics provides only part of the picture in multiclass prediction. This section delves into how model performance varies across the six prediction classes, revealing a clear hierarchy in outcome predictability that has important implications for practical deployment. By examining performance at the class level, we can identify which types of predictions are most reliable and understand the underlying factors that drive predictability differences.

Table 6.2 breaks down the classification metrics for each outcome type using the best-performing model (LightGBM optimized for Macro-F1). This granular analysis reveals substantial variation in predictability across different fight outcomes, providing crucial insights for users who need to understand when model predictions are most trustworthy.

Table 6.2: Per-class F1 scores for the best multiclass model (LightGBM optimized for Macro-F1). The results reveal a clear hierarchy in predictability, with decisions being most predictable and submissions most challenging.

Outcome Class	Fighter 1 F1	Fighter 2 F1
Decision	0.411	0.389
KO/TKO	0.240	0.249
Submission	0.111	0.087

The results reveal a clear hierarchy in predictability that reflects both the statistical properties of the data and the inherent nature of different victory methods: **Decisions easiest → KO/TKO → Submissions hardest**. A clear predictability hierarchy emerges from the per-class results. Decisions are the most predictable

outcome (F1 $\approx$ 0.40), benefiting from a larger sample size and more distinct fighter profiles. KO/TKO outcomes are moderately predictable (F1 $\approx$ 0.24), as strike-based features provide useful signals that are nonetheless limited by the inherent volatility of knockouts. Submissions prove to be the most challenging class to predict (F1 $\approx$ 0.10), a consequence of their relative rarity and opportunistic nature. Across all methods, the model maintains fighter symmetry, with similar F1 scores for Fighter 1 and Fighter 2 indicating no systematic bias. This hierarchy provides valuable guidance for practical applications, helping users understand which predictions carry the most confidence.

6.4.1 Decision Outcomes (F1 $\approx$ 0.40)

Decisions emerge as the most predictable outcome, achieving F1-scores around 0.40 for both fighters. This superior performance stems from multiple factors working in concert. The larger sample size, with decisions comprising approximately 56% of all fights, provides more training examples for the model to learn from. Importantly, decisions represent the default outcome when neither fighter achieves a finish, making them somewhat easier to predict by process of elimination.

The relatively high predictability of decisions has important practical implications. For betting applications, decision predictions can form the foundation of a conservative strategy. For strategic preparation, knowing a fight is likely to go the distance allows teams to focus on cardio conditioning and point-scoring techniques rather than finish-seeking strategies.

6.4.2 KO/TKO Outcomes (F1 $\approx$ 0.24)

Knockout predictions show moderate success with F1-scores around 0.24. The model demonstrates ability to capture signals related to finishing power and defensive vulnerabilities, though with less certainty than for decisions. Several factors might contribute to both the partial success and limitations in knockout prediction. Strike-based features in our dataset, such as significant strikes landed and absorbed per minute, provide clear signals about a fighter's offensive output and defensive gaps. Large disparities in historical knockout rates between opponents could create strong predictive signals. However, the inherent volatility of knockouts, where a single clean

punch can end a fight regardless of who was winning, introduces irreducible randomness that limits achievable accuracy.

This moderate predictability suggests that while the model can identify fights with elevated knockout potential, predicting the exact occurrence remains challenging. Users should interpret knockout predictions as risk assessments rather than definitive forecasts, understanding that even high-probability knockout predictions carry substantial uncertainty.

6.4.3 Submission Outcomes (F1 $\approx$ 0.10)

Submissions prove most challenging, with F1-scores around 0.10 reflecting the difficulty of predicting these outcomes. This performance level, while low in absolute terms, still represents meaningful signal extraction given the rarity and complexity of submissions. The prediction difficulty arises from several compounding factors. With only approximately 16% of fights ending in submission, the model has limited training examples, particularly for specific fighter matchups. Many submissions arise from brief scrambles, momentary lapses in defense, or opponent mistakes that are essentially impossible to predict from historical statistics. Our current feature set may inadequately capture grappling nuances such as guard retention, submission defense patterns, or the chess-like progression of positional grappling. Finally, submission outcomes show high variance even for the same fighter, as a grappling specialist might achieve multiple submissions or none at all depending on opponent styles and fight dynamics.

The low predictability of submissions suggests they should be treated as high-uncertainty events in practical applications. However, the model's ability to identify any signal in this challenging domain provides value for users who understand these limitations and can incorporate the uncertainty into their decision-making processes.

The symmetry in F1 scores between Fighter 1 and Fighter 2 across all outcome types indicates that our model maintains fairness and avoids systematic bias toward either fighter position, an important property for practical deployment. This symmetry confirms that the model treats both fighters equally, learning patterns based on their attributes rather than their position in the dataset.

The clear performance hierarchy across outcome types provides essential context for interpreting model predictions. Users can calibrate their confidence based on

the predicted outcome type, placing more trust in decision predictions while treating submission predictions as lower-confidence indicators of potential outcomes.

6.5 Comparison with Binary Classification

A critical question for practical deployment is whether the added complexity of multiclass prediction compromises basic win/loss accuracy. This section demonstrates that multiclass models not only maintain strong binary performance but offer an exceptional trade-off between granularity and accuracy. The analysis reveals that users need not choose between simple and detailed predictions, as multiclass models effectively serve both purposes.

6.5.1 Binary Performance from Multiclass Models

To evaluate binary performance, we aggregate the predicted probabilities for all Fighter 1 victory methods (Decision, KO/TKO, Submission) to derive binary win probabilities. This aggregation approach allows us to assess whether the multiclass model can serve dual purposes: providing detailed method predictions when needed while maintaining competitive win/loss predictions for simpler applications.

The aggregated predictions yield impressive binary performance metrics. The best multiclass model (LightGBM optimized for Macro-F1) achieves 67.9% binary accuracy, nearly matching the dedicated binary models. The Binary AUC ranges from 0.723 to 0.736, comparable to our dedicated binary models' performance. Binary Brier Scores of 0.214-0.215 are only slightly higher than the best binary models, indicating maintained calibration quality.

For context, the best dedicated binary models in our experiments achieved:

- LightGBM: 67.3% accuracy (historical best: 69.7%)
- XGBoost: 67.0% accuracy (historical best: 70.59%)

The multiclass models thus retain approximately 96% of the best historical binary performance while adding six-fold granularity, an excellent trade-off for many applications.

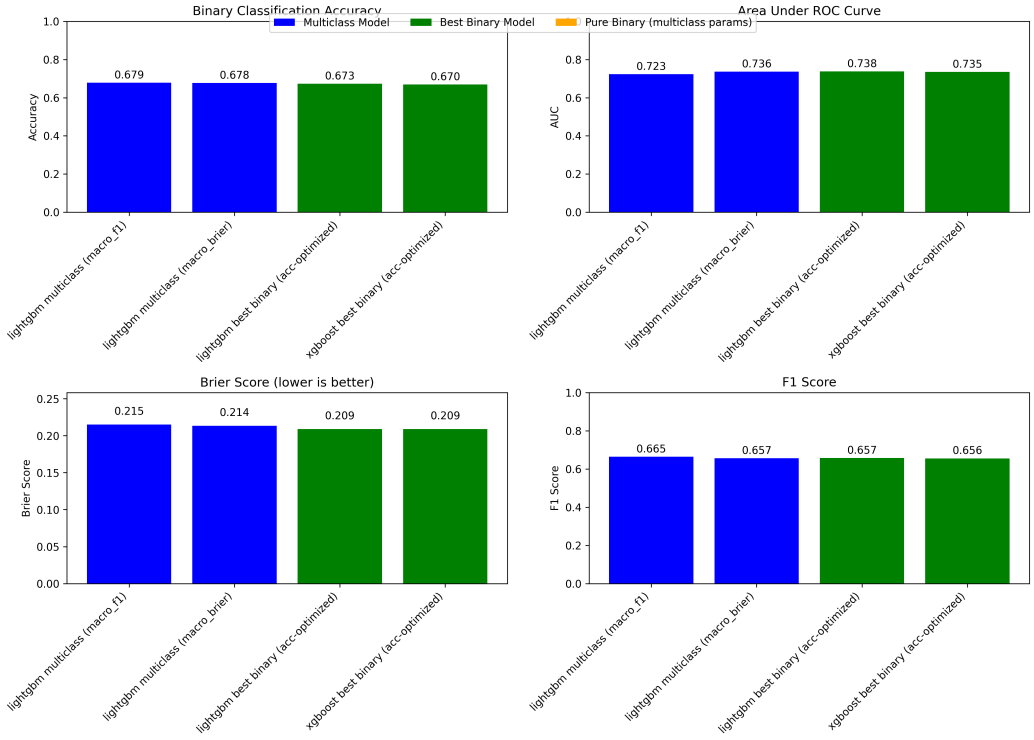


Figure 6.1: Performance comparison between dedicated binary models and multiclass models evaluated on the binary task. The multiclass-derived binary predictions maintain strong performance while providing six-fold more granular predictions.

6.5.2 Feature Importance Patterns

The feature importance analysis reveals both similarities and differences compared to binary classification, providing insights into how the model adapts to the increased complexity of multiclass prediction. Understanding these patterns helps explain why multiclass models maintain strong binary performance while successfully predicting specific methods.

External signals (betting odds and Google Trends) maintain their importance in the multiclass setting, collectively contributing approximately 20% of total feature importance. This consistency suggests that market wisdom and public sentiment provide value regardless of prediction granularity. The balanced approach observed in binary classification persists, with external signals complementing rather than replacing traditional fighter statistics. Importantly, the contribution of external signals remains stable across different outcome types, indicating their general rather than method-specific value.

This stability in feature importance patterns partially explains the strong binary performance of multiclass models. The same fundamental signals that predict winners in binary classification continue to drive predictions in the multiclass setting, with additional nuanced patterns emerging to differentiate between specific victory methods. This architectural similarity ensures that multiclass models retain the predictive foundations of binary models while adding layers of sophistication.

6.5.3 Practical Trade-offs

The choice between binary and multiclass models involves clear considerations that depend on specific use cases and constraints. Understanding these trade-offs helps practitioners select the appropriate model architecture for their applications.

Binary models remain the optimal choice when maximum accuracy is paramount, computational resources are limited, or when only win/loss predictions are needed. The marginal accuracy advantage and simpler deployment make them suitable for applications focused solely on predicting winners. Additionally, binary models require less computational resources for training and inference, making them attractive for resource-constrained environments.

Multiclass models excel when method-specific insights provide value. Applications

benefiting from detailed predictions include betting on multiple markets simultaneously, strategic fight preparation requiring understanding of likely victory paths, commentary and analysis seeking deeper insights, and unified deployment scenarios where maintaining a single model simplifies system architecture. The ability to serve both simple and detailed prediction needs makes multiclass models particularly attractive for comprehensive MMA analytics platforms.

The key insight is the exceptional information-to-accuracy trade-off: multiclass models provide 3× more information with minimal accuracy sacrifice on the binary task. This makes them highly attractive for applications that can leverage the additional granularity. For many practical deployments, the small accuracy sacrifice is more than compensated by the richness of information provided.

This analysis demonstrates that multiclass models represent a mature and practical approach to MMA prediction, capable of serving diverse application needs while maintaining strong performance across multiple evaluation criteria. The next sections explore additional aspects of multiclass performance, including temporal dynamics and feature importance patterns.

6.6 Temporal Dynamics and Window Optimization

The selection of training data recency profoundly impacts model performance in the multiclass setting. This section examines how different models respond to varying temporal windows and what these patterns reveal about the underlying dynamics of MMA prediction. Understanding these temporal patterns is crucial for maintaining model performance over time and adapting to the evolving nature of the sport.

Figure 6.2 illustrates the relationship between temporal window size and model performance across all optimization trials. The visualization reveals distinct patterns that differ markedly between model types and optimization objectives. We observe that performance responses to temporal window changes are non-monotonic, with clear optimal ranges emerging for each model type.

The varied optimal windows (4.5-7.0 years) reveal sophisticated model-specific dynamics that warrant detailed examination. These differences are not random variations but reflect fundamental differences in how each algorithm processes temporal patterns in the context of increased prediction complexity.

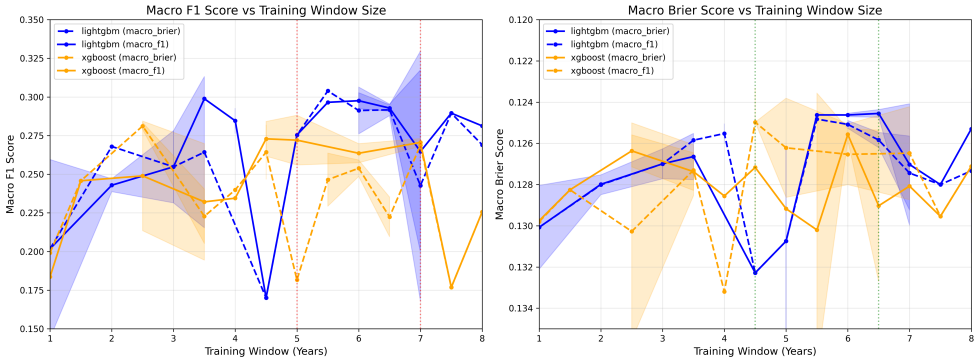


Figure 6.2: Impact of temporal window size on multiclass performance. The plot shows Macro-F1 and Multiclass Brier scores as functions of training years, with confidence bands representing variance across trials. Performance shows model-specific optima: XGBoost favors 4.5-5.0 years while LightGBM prefers 6.5-7.0 years.

Model architecture profoundly influences temporal preferences. XGBoost's consistent preference for shorter windows (4.5-5.0 years) suggests a focus on capturing recent trends and current fighter form. This may reflect XGBoost's depth-first tree construction, which creates more specialized decision paths that benefit from temporal consistency. In contrast, LightGBM's preference for longer windows (6.5-7.0 years) indicates a strategy of leveraging broader historical patterns. The leaf-wise growth may better handle the heterogeneity introduced by longer time spans, finding value in the increased data volume despite potential concept drift.

Optimization objectives also drive temporal variation. Models optimized for Brier score tend toward slightly different windows than F1-optimized models, indicating that the temporal relevance of data varies depending on whether we prioritize classification accuracy or probability calibration. This suggests that recent data may be more valuable for accurate classification, while longer histories help calibrate probability estimates across rare events. The difference is particularly pronounced for submission predictions, where longer histories provide more examples of these rare events.

The balance between data sufficiency and recency becomes particularly critical in the multiclass setting. With six imbalanced classes, models must navigate competing demands: having enough historical examples to learn patterns for rare outcomes like submissions, while maintaining temporal relevance as the sport evolves. Each model type resolves this trade-off differently, leading to the observed variation in optimal windows. XGBoost appears to prioritize recency, accepting fewer examples of rare

events in exchange for temporal consistency. LightGBM takes the opposite approach, valuing the statistical power of larger datasets even at the cost of including potentially outdated patterns.

These temporal dynamics have important implications for model deployment and maintenance. Organizations using XGBoost models should plan for more frequent retraining to maintain optimal performance, while LightGBM deployments can operate with longer update cycles. Understanding these patterns also helps explain performance degradation over time, a topic explored in detail in Chapter 7.

6.7 Method-Specific Feature Importance

Moving beyond aggregate performance metrics, this section examines which features drive predictions for each specific outcome type. Through SHAP (SHapley Additive exPlanations) analysis, we uncover the nuanced patterns in how different feature categories contribute to various victory method predictions. This granular analysis provides insights into the decision-making process of our models and helps validate that the learned patterns align with domain expertise.

6.7.1 Feature Categories and Analysis Framework

Our analysis leverages the comprehensive metadata system to organize features into semantically meaningful categories. This categorization enables us to understand not just which individual features matter, but what types of information drive different predictions. By examining category-level feature importance, we can derive insights that are both more interpretable and more actionable than individual feature analysis alone.

The feature categories encompass several distinct types of information. Betting Odds comprise 11 features capturing market-derived probabilities and implied win percentages, representing the collective wisdom of the betting market. Google Trends include 57 features measuring search volume data and public interest dynamics, potentially capturing momentum, media attention, and fan sentiment. Physical Attributes cover fundamental characteristics like height, reach, and weight class indicators. Performance Statistics aggregate historical fight outcomes including strike

accuracy, takedown success rates, and finishing patterns. Timing Features capture career dynamics through metrics like days since last fight and total career duration.

This categorization allows us to move beyond examining individual features to understanding what types of information are most valuable for different predictions. For instance, knowing that "striking features" are important is more actionable than knowing that a specific striking metric matters, as it suggests a whole class of related features that could be collected or engineered.

6.7.2 Overall Feature Importance by Category

Figure 6.3 presents the aggregate feature importance across all six prediction classes, revealing how different information sources contribute to the model's decision-making process. We observe a clear hierarchy in feature importance that provides insights into the fundamental drivers of fight outcome predictions.

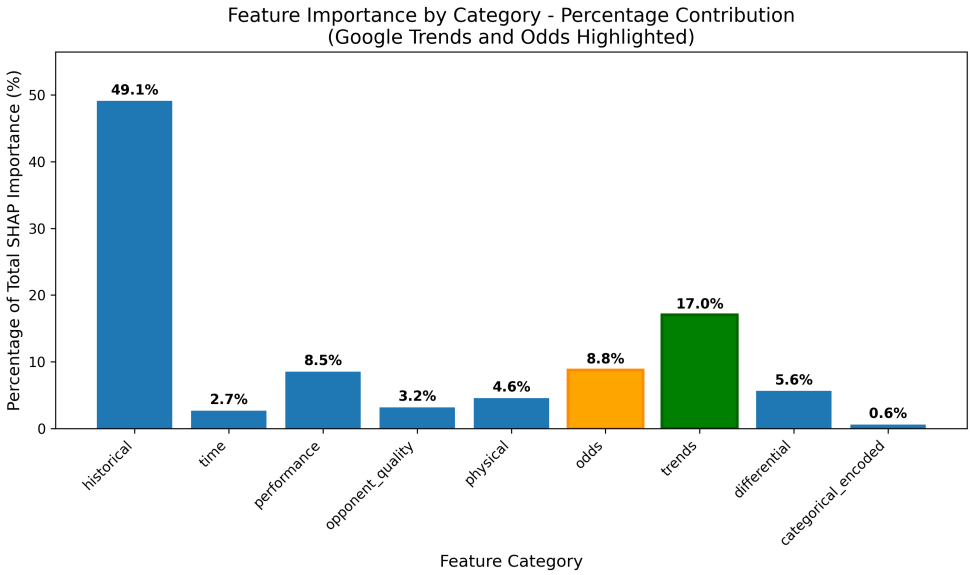


Figure 6.3: SHAP feature importance by category across all six prediction classes. External signals (Betting Odds and Google Trends) contribute approximately 20% of total importance, providing valuable complementary information to traditional fighter statistics.

The importance distribution reveals a balanced and intuitive pattern. Traditional fighter statistics form the foundation of predictions, contributing approximately 80% of total predictive power through historical performance metrics, physical attributes, and timing features. This dominance of objective fighter data confirms that past

performance remains the strongest predictor of future outcomes, aligning with the fundamental principle that fighter skill and historical patterns drive fight results.

External signals provide meaningful refinements, with betting odds and Google Trends collectively contributing about 20% of model importance. Rather than overwhelming the model, these features add layers of information not captured in raw statistics. Market assessments encode collective wisdom about matchup dynamics, training camp reports, and factors difficult to quantify in traditional statistics. Public sentiment captured through search patterns may reflect momentum shifts, media narratives, or fan awareness of factors not yet reflected in fight records.

This 80/20 split represents a synergistic relationship where objective fighter data provides the foundation while subjective market assessments add valuable refinements. The model effectively combines these complementary information sources to achieve superior predictions, demonstrating that optimal performance comes from intelligent feature combination rather than relying on any single information source.

6.7.3 Method-Specific Patterns

Figure 6.4 breaks down feature importance by specific outcome type, revealing whether certain features become more or less important for predicting particular victory methods. The analysis provides crucial insights into whether the model learns method-specific patterns or relies on general fight outcome signals.

The analysis reveals consistent patterns across outcome types, with subtle but interpretable variations that provide insights into how different victory methods are predicted.

Decision Outcomes

For fights ending in decisions, external signals contribute 20.5% of total importance (Google Trends: 12.7%, Betting Odds: 7.8%). This slightly elevated importance may reflect that decisions are often easier for markets to predict, as they typically involve known stylistic matchups between durable fighters. The higher betting odds contribution (7.8%) for decisions could indicate that markets are particularly good at identifying fights likely to go the distance, possibly because decision-prone fighters have established patterns that markets efficiently recognize.

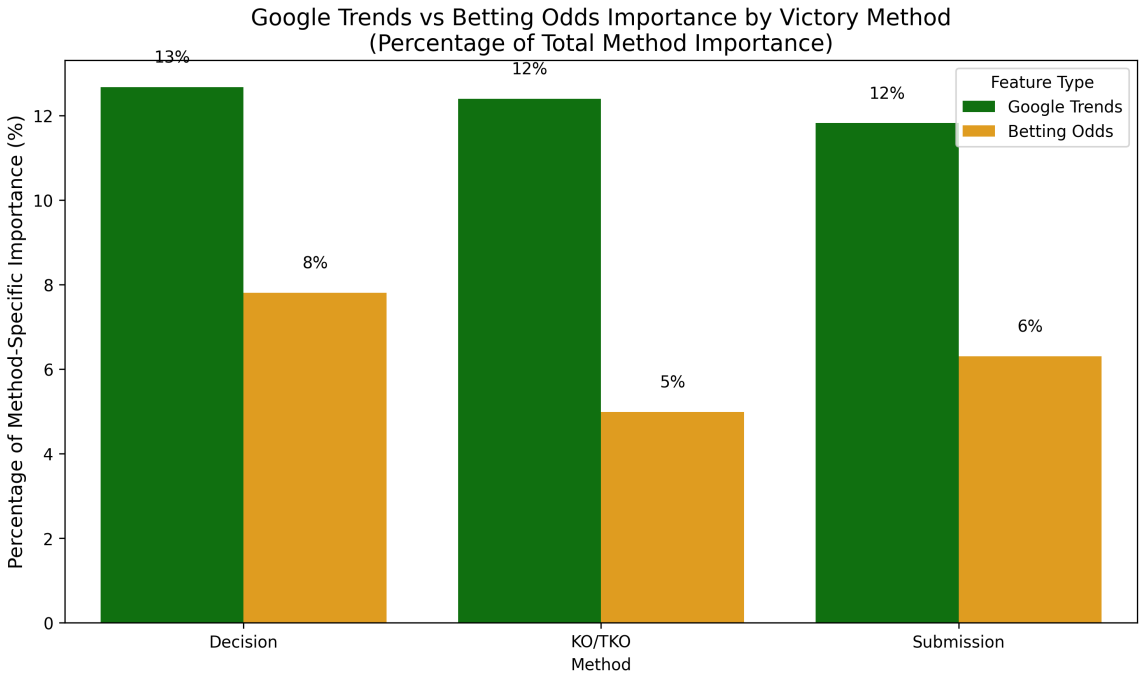


Figure 6.4: Feature importance percentages segmented by method of victory. External signals show consistent importance across all outcome types, with slight variations: Google Trends contribute 11.8-12.7% while Betting Odds contribute 5.0-7.8% depending on the method.

KO/TKO Outcomes

Knockout predictions show the lowest reliance on external signals at 17.4% total (Google Trends: 12.4%, Betting Odds: 5.0%). The reduced importance of betting odds (5.0%) suggests that markets may struggle to predict the inherent volatility of knockouts. Google Trends maintain their importance, possibly capturing fighter momentum or training camp reports about improved striking. The lower overall external signal contribution implies that knockout predictions rely more heavily on objective striking statistics and power metrics.

Submission Outcomes

Submission predictions fall between the extremes with 18.1% external signal importance (Google Trends: 11.8%, Betting Odds: 6.3%). Given the rarity and technical nature of submissions, one might expect very different feature importance patterns. The consistency suggests that external signals provide general fight outcome information rather than method-specific insights. This pattern indicates that while markets and public sentiment can identify likely winners, they provide limited additional information about specific submission threats.

The consistent patterns across methods (17-21% for external signals) indicate these features provide broad predictive value about fight outcomes rather than specialized insights about specific finishing methods. This pattern suggests that market wisdom and public sentiment capture general fight dynamics that translate across all potential outcomes, reinforcing the value of including these signals in a comprehensive prediction system.

6.7.4 Interpretation of External Signal Importance

The stable 20% contribution of external signals across all outcome types provides several important insights for understanding their role in MMA prediction. This consistency suggests fundamental properties of how markets and public sentiment relate to fight outcomes.

Market efficiency manifests through betting odds that encode collective wisdom about matchups, incorporating information that may not be fully captured by historical statistics alone. This includes factors like stylistic matchups, training camp

reports, and recent form that are difficult to quantify but influence fight outcomes. The consistent importance across outcome types suggests markets efficiently process available information to predict overall fight dynamics rather than specific methods.

Google Trends serve as current form indicators, potentially capturing momentum shifts, injury concerns, or training camp developments not yet reflected in historical fight data. The search patterns may also reflect media narratives and public perception that, while not directly causal, correlate with fight outcomes. The stability of their contribution suggests these signals provide consistent value regardless of how fights end.

The complementary nature of these signals cannot be overstated. Rather than replacing traditional analysis, external signals enhance it by adding dimensions of information that pure statistics cannot capture. Their 20% contribution represents a meaningful but not dominant enhancement to model performance. This balance ensures that models remain grounded in objective fighter capabilities while benefiting from the collective intelligence of markets and crowds.

The consistent value across outcome types suggests these signals capture fundamental aspects of fight dynamics rather than method-specific patterns. This broad applicability makes them valuable additions to any MMA prediction system, regardless of the specific outcomes being predicted. Organizations implementing these models can confidently include external signals knowing they provide stable value across all prediction scenarios.

6.8 Error Analysis and Confusion Patterns

Understanding model failures provides crucial insights for practical deployment and future improvements. This section analyzes the systematic patterns in model errors through detailed examination of the confusion matrix, revealing both the strengths and limitations of multiclass MMA prediction. By understanding where and why the model fails, users can better calibrate their confidence in different types of predictions and developers can identify areas for improvement.

Figure 6.5 presents the normalized confusion matrix for our best-performing model. The visualization reveals clear patterns in how the model succeeds and fails, with important implications for real-world applications. We observe distinct error patterns

that inform both model interpretation and practical deployment strategies.

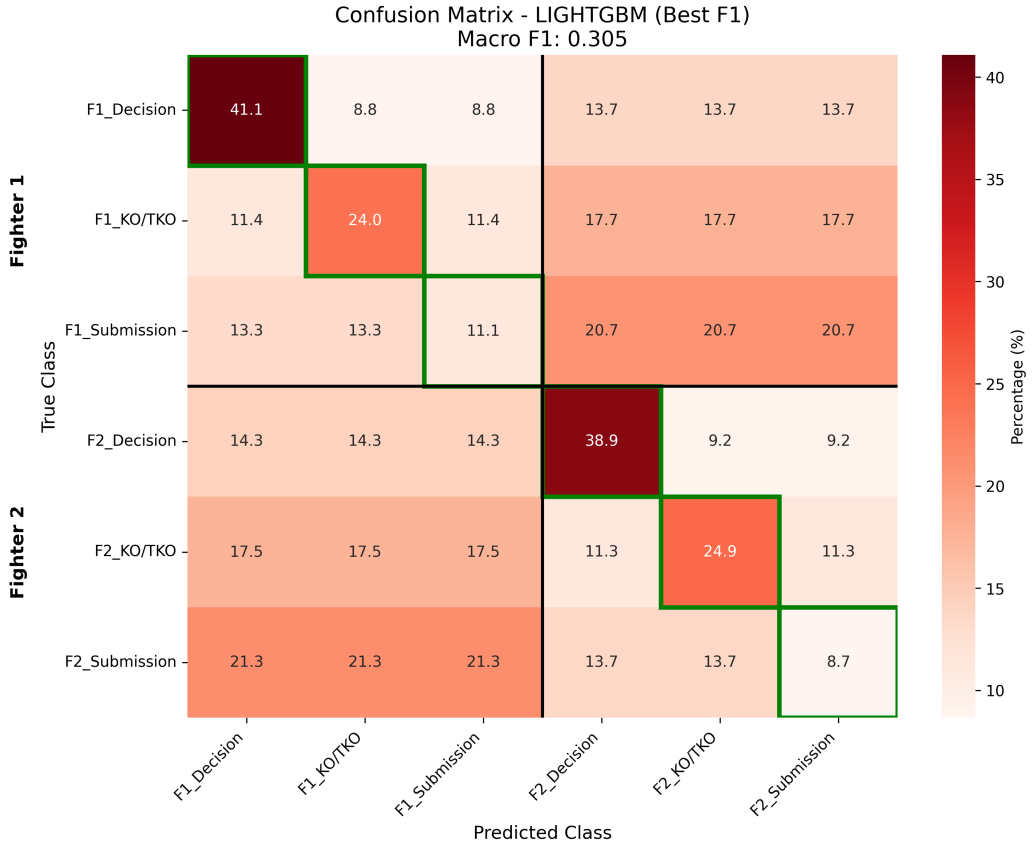


Figure 6.5: Normalized confusion matrix for the best multiclass model (LightGBM optimized for Macro-F1). Darker cells indicate higher confusion rates. The matrix reveals that winner prediction errors are less common than method prediction errors, with the highest confusion occurring between different methods for the same winner.

Analysis of the confusion matrix reveals a clear hierarchy of prediction difficulty that has important implications for model deployment and interpretation. The patterns visible in this matrix tell a story about what aspects of fight prediction are fundamentally tractable versus those that remain challenging even with sophisticated models.

6.8.1 Winner vs Method Confusion

The confusion patterns reveal fundamental insights about what aspects of fight prediction are tractable versus challenging. As shown in Figure 6.5, the model rarely confuses Fighter 1 victories with Fighter 2 victories, as evidenced by the minimal values in off-diagonal blocks of the matrix. This indicates that the binary signal of

who wins remains strong even in the multiclass setting. The features that predict winners, including skill differentials, recent form, and physical advantages, translate well regardless of how specifically the fight ends.

In contrast, most confusion occurs within the same fighter's victory methods. For example, when the model predicts Fighter 1 will win by decision but Fighter 1 actually wins by KO/TKO, this represents a correct winner prediction but incorrect method prediction. This pattern dominates the error landscape, suggesting that while we can reliably predict who will win, predicting exactly how remains challenging. The within-fighter confusion reflects the inherent uncertainty in fight dynamics, where small moments can shift the outcome from one method to another.

The decision-finish boundary emerges as particularly problematic. Many fights predicted as decisions actually end in late stoppages, often in the third round when accumulated damage finally overwhelms a fighter. Conversely, some predicted finishes go to the judges when durable fighters survive early adversity. This boundary is inherently fuzzy because a fight ending with 10 seconds remaining is functionally similar to one that goes to decision, yet they belong to different classes. This fuzzy boundary represents an inherent limitation of the classification approach rather than a model failure.

6.8.2 Confidence Calibration and Deployment Considerations

Despite moderate absolute accuracy, the model demonstrates several properties that ensure practical utility in real-world applications. Understanding these properties helps users leverage model outputs effectively despite the inherent challenges of multiclass prediction.

The strong binary signal preservation means that users can rely on the model for the fundamental prediction of who wins, even if the specific method prediction carries more uncertainty. This property allows graceful degradation in application design. When method predictions are uncertain, users can fall back to well-calibrated winner predictions, ensuring that the model provides value even in challenging prediction scenarios.

Even with lower absolute accuracy for specific methods, the model effectively ranks the relative likelihood of different victory methods. A prediction of 40% decision, 35% KO/TKO, and 25% submission meaningfully indicates that a decision is most

likely, even if the specific outcome remains uncertain. This ranking ability provides value for applications that need to assess relative probabilities rather than make hard classifications. For instance, a betting application can use these rankings to identify value across multiple betting markets simultaneously.

The low Brier scores across all configurations indicate that predicted probabilities align well with empirical frequencies. When the model assigns a 30% probability to a specific outcome, that outcome indeed occurs approximately 30% of the time across many predictions. This calibration property is essential for decision-making applications, particularly in betting contexts where expected value calculations depend on accurate probability estimates. Well-calibrated probabilities ensure that users can make optimal decisions based on model outputs.

For practical deployment, these properties suggest a nuanced approach to using model outputs. Rather than treating predictions as definitive classifications, users should leverage the full probability distribution to make informed decisions that account for prediction uncertainty while capitalizing on the model's strengths in winner prediction and relative method ranking. Applications should be designed to gracefully handle the inherent uncertainty in method predictions while fully exploiting the strong winner predictions and well-calibrated probability estimates.

These insights into error patterns and model properties provide essential guidance for both current users and future development efforts. By understanding where the model excels and where it struggles, we can design applications that maximize value while honestly representing prediction uncertainty.

6.9 Summary and Key Findings

This chapter has comprehensively demonstrated that multiclass MMA prediction, while more challenging than binary classification, provides substantial value through granular insights and maintains strong practical performance. Our analysis yields several key findings that inform both the current state of MMA prediction and future research directions.

The achievement of meaningful multiclass performance stands as the primary success. The best model (LightGBM optimized for Macro-F1) achieved a 0.305 F1-score, representing nearly 1.8× improvement over random baseline predictions. With strong

discrimination (0.783 Macro-AUC) across all six classes, the model successfully captures predictive patterns despite the inherent complexity of predicting both winner and method. This performance level demonstrates that the additional complexity of multiclass prediction does not prevent the extraction of meaningful signals from the data.

Feature importance analysis reveals a balanced and intuitive pattern. External signals comprising betting odds and Google Trends contribute approximately 20% of predictive power, providing valuable market wisdom and sentiment information. This contribution complements rather than dominates the 80% contribution from traditional fighter statistics. The synergistic relationship between objective fighter data and subjective market assessments creates a robust prediction framework that leverages multiple information sources effectively.

Multiclass models retain strong binary classification capability. When evaluated on the simple win/loss task through probability aggregation, multiclass models achieve 67.9% accuracy, approximately 96% of the best dedicated binary model performance. This minimal accuracy trade-off for six-fold increased granularity represents a practical value proposition for many applications. Organizations need not maintain separate models for different prediction tasks, as multiclass models effectively serve both detailed and simple prediction needs.

A clear predictability hierarchy emerges across outcome types: **Decisions easiest → KO/TKO → Submissions hardest**. Decisions prove most predictable with F1-scores around 0.40, reflecting both their frequency in the dataset and the identifiable patterns of fighters who typically go to decision. Knockouts show moderate predictability (F1≈0.24), with strike-based features providing signals tempered by inherent volatility. Submissions remain most challenging (F1≈0.10), limited by their rarity and opportunistic nature. This hierarchy provides essential context for interpreting model predictions and calibrating confidence appropriately.

Model-specific temporal dynamics reveal sophisticated patterns in how different algorithms process historical data. XGBoost consistently prefers shorter windows of 4.5-5.0 years, focusing on recent trends and current form. LightGBM favors longer windows of 6.5-7.0 years, leveraging broader historical patterns. These differences reflect fundamental variations in how model architectures handle temporal heterogeneity and the trade-off between data recency and sufficiency. Understanding these

patterns guides deployment decisions and maintenance schedules.

The analysis of error patterns through confusion matrices reveals that winner prediction remains robust while method prediction drives most errors. The model rarely confuses victories between fighters but frequently misclassifies the specific method of victory. This pattern, combined with well-calibrated probabilities, suggests deployment strategies should leverage the strong winner predictions while using method probabilities for relative ranking rather than absolute classification. Applications designed with this understanding can provide maximum value while honestly representing prediction uncertainty.

These findings demonstrate that multiclass MMA prediction transcends mere technical feasibility to provide genuine practical value. By effectively combining traditional sports analytics with market wisdom and public sentiment, the models achieve meaningful performance that can inform strategic preparation, enhance betting strategies, and deepen analytical understanding of the sport. The balanced approach, where external signals refine rather than replace fighter statistics, points toward future prediction systems that synthesize multiple information sources to achieve superior performance.

Looking forward, the insights from this chapter lay the groundwork for exploring how these predictions perform in real-world deployment scenarios. The strong performance, clear interpretability, and practical utility of multiclass models position them as valuable tools for the MMA analytics ecosystem. The next chapter examines post-fight model drift, investigating how predictions degrade over time and what this reveals about the dynamic nature of MMA competition.

Chapter 7

Postfight Analysis: Quantifying Data Drift in MMA

7.1 Introduction to Drift in MMA: Understanding Temporal Instability

MMA constantly evolves through changing fighter strategies, training methods, and rules. This evolution creates data drift that challenges predictive modeling.

This chapter presents a detailed postfight analysis of data drift within the UFC dataset compiled for this thesis. The primary objective is to demonstrate that the walk-forward validation framework established in Chapter 4 is necessary for temporal stability. The analysis focuses particularly on the COVID-19 pandemic period (March 2020 to April 2021), which represented an unprecedented disruption to the sport's normal operations. The analysis is structured to answer several critical research questions.

First, we seek to identify at which specific points in recent UFC history significant data drift has occurred. Second, we aim to quantify the impact of this drift on the performance of our predictive models, as measured by metrics like Log Loss, Brier Score, and Accuracy. Third, we investigate which specific fighter statistics and attributes serve as the primary drivers of this drift. Fourth, we examine whether the observed drift can be correlated with known, real-world events in the history of MMA. Fifth, we analyze how the unique circumstances of the COVID-19 pandemic affected fighter performance patterns. Finally, we compare how static machine learning models

perform relative to continuously adaptive systems like betting markets.

To answer these questions, this chapter systematically examines drift through three analytical lenses. We begin with model performance drift, tracking the temporal degradation of key model performance indicators. We then analyze feature distribution drift, statistically measuring the shift in the distributions of individual features over time using the Kolmogorov-Smirnov (KS) test and Population Stability Index (PSI). Finally, we investigate concept drift by examining changes in the relationships between features and fight outcomes through analyzing the evolution of SHAP (SHapley Additive exPlanations) values.

The findings detailed here not only illuminate the specific temporal dynamics of the UFC but also provide a crucial foundation for understanding the limitations of static models and the necessity of adaptive learning systems in sports analytics.

7.2 Methodology for Temporal Analysis

The methodology for this drift analysis follows a rigorously defined analytical framework that ensures a robust and reproducible investigation into the temporal dynamics of the dataset. The initial reference model for this analysis was trained on five years of data ending on January 1, 2017.

For this analysis, we intentionally employ a different, simpler model than the optimized versions in Chapters 5 and 6. This model is deliberately **static**: it is trained only once on data up to January 2017 and is never retrained. This "deploy-and-forget" approach simulates a common production scenario and provides a fixed, consistent baseline, making it possible to isolate and quantify the effects of temporal drift. Consequently, its performance is expected to be lower than adaptively trained models, as the primary goal here is to measure degradation, not to achieve peak accuracy.

7.2.1 Drift Detection Framework

The core of our analysis is a temporal, event-based framework that processes data in chronological order to simulate a real-world model deployment scenario. The data is segmented and analyzed using a sliding window approach.

The key parameters of this framework include temporal windows, minimum fight requirements, and reference strategies. The analysis is conducted using a **3-month**

sliding window, a duration selected as an optimal trade-off that captures a sufficient number of events for statistical significance while remaining sensitive to relatively rapid changes in the sport. Each 3-month window is required to contain a minimum of **50 fights** to maintain statistical validity, with windows failing to meet this threshold excluded from the analysis. The analysis employs a **sliding reference window**, comparing each new 3-month block of data to the immediately preceding 6 months, an approach highly effective for detecting recent and ongoing changes in the data's statistical properties.

7.2.2 Metrics and Statistical Tests

Drift is quantified using a suite of metrics and statistical tests applied to model performance, feature distributions, and feature importance.

For performance drift, we monitor the Z-score of the **Log Loss**, a sensitive metric that penalizes both inaccurate and overconfident predictions. A Z-score exceeding a threshold of **2.0** signals a "Warning" for potential drift, while a score above **3.0** indicates a "Critical" drift alert.

To detect changes in feature distributions, we employ two industry-standard tests. The **Kolmogorov-Smirnov (KS) Test** uses a p-value threshold of **0.01**, where a p-value below this threshold indicates a statistically significant difference between the feature's distribution in the current window and the reference window. The **Population Stability Index (PSI)** uses a threshold of **0.2**, where a PSI value above this threshold is considered to represent a major shift in the feature's distribution.

Changes in the underlying relationships between features and outcomes are analyzed by tracking the evolution of SHAP values for each feature, comparing pre-COVID (2016-2019) and COVID period (2020-2022) data.

7.3 Empirical Results of Drift Analysis

This section details the empirical findings, presenting the detected drift in model performance, feature distributions, and feature importance (concept drift).

7.3.1 Model Performance Drift: COVID-19 as a Catalyst

The primary indicator of instability in our predictive model is its performance over time. Analysis across 34 time windows revealed an overall model performance of **0.640 ± 0.040** (mean ± standard deviation) in accuracy and **0.635 ± 0.026** (mean ± standard deviation) in log loss. This level of performance is consistent with our experimental design, reflecting the intentional use of a static model to measure the impact of drift over time. Critically, the log loss Z-score analysis reveals significant performance degradation over time. This finding underscores the necessity of regular model retraining and directly validates the event-based walk-forward framework used in this thesis. The Z-score exceeded the 'Warning' threshold of 2.0 in multiple windows and the 'Critical' threshold of 3.0 at the height of the pandemic's disruption, as visualized in Figure 7.1.

This performance drift is particularly striking when contrasted with the stability of betting markets (Section 7.3.4). While a static model trained on historical data suffers progressive degradation, betting markets maintain consistent accuracy by continuously adapting to new information, highlighting the fundamental limitation of deploy-and-forget machine learning models in dynamic sports environments.

Figure 7.1 illustrates the model's performance metrics from 2017 to 2023, with particular emphasis on the COVID-19 period (March 2020 to April 2021). The shaded region highlights this unprecedented period in UFC history, during which the sport underwent dramatic operational changes.

The COVID-19 period introduced unique challenges to MMA prediction. Empty arena events fundamentally altered the fight atmosphere, potentially affecting fighter psychology and performance due to the absence of crowd energy. The creation of UFC Fight Island in Abu Dhabi introduced controlled but artificial conditions through a quarantine bubble system. Travel restrictions and gym closures forced fighters to adapt their preparation methods, with many relying on remote coaching and limited sparring partners, disrupting traditional training camps. Additionally, many fighters experienced longer periods between fights due to event cancellations and rescheduling, leading to extended layoffs that affected their competitive rhythm.

These environmental factors contributed to increased performance variability, though the model maintained its overall predictive capability throughout this challenging period.

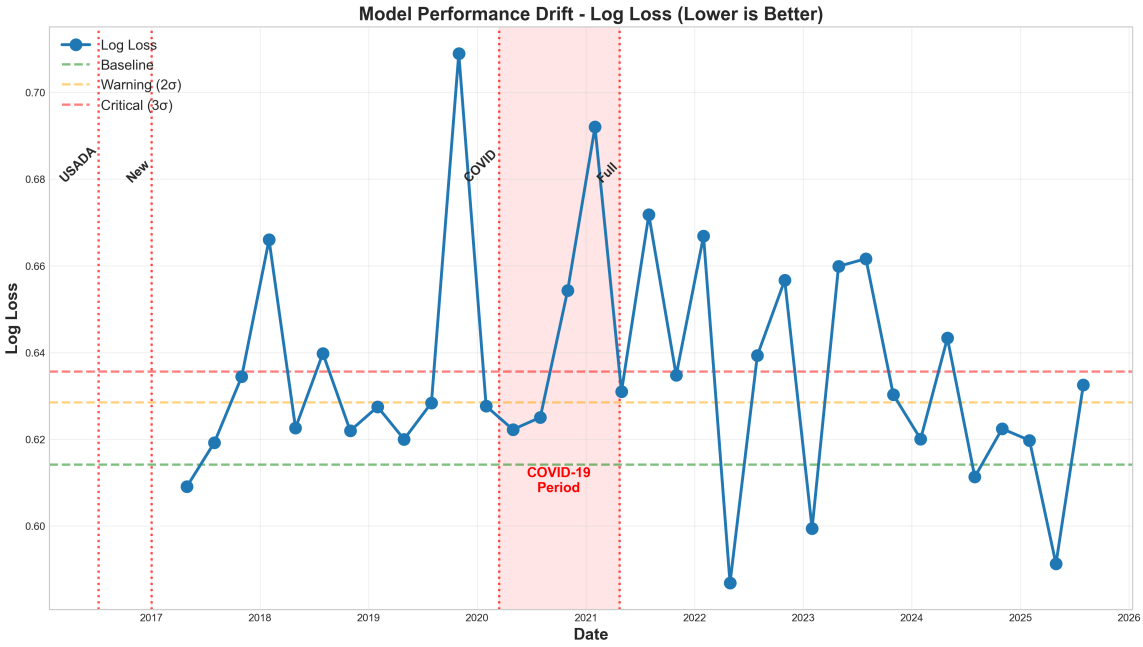


Figure 7.1: The model's Log Loss over time, calculated in 3-month windows. The shaded COVID-19 period (March 2020 - April 2021) shows notable performance variations. Red dotted lines indicate known UFC events for correlational analysis.

7.3.2 Feature Distribution Drift: Quantifying the Evolution

The performance degradation detailed previously is a direct result of underlying shifts in the data distributions. To diagnose these shifts, the feature distribution analysis examined **335 features** across 35 time windows, revealing an average drift rate of **4.8%** with a maximum of **49.6%**. This comprehensive analysis provides insights into which aspects of fighter performance underwent the most significant changes.

Figure 7.2 presents the Population Stability Index (PSI) values for the top drifted features, where each 3-month window is compared against the preceding 6-month reference period. In this visualization, darker red colors indicate higher PSI values, representing greater drift. Values above 0.2 (shown in the darkest red) indicate material drift that could significantly impact model performance. The heatmap reveals distinct patterns of drift, with certain features showing pronounced changes during specific periods.

The Kolmogorov-Smirnov test analysis provides complementary insights into feature distribution changes. Figure 7.3 presents the KS test p-values for the top 20 features exhibiting significant distributional shifts. In this heatmap, red indicates

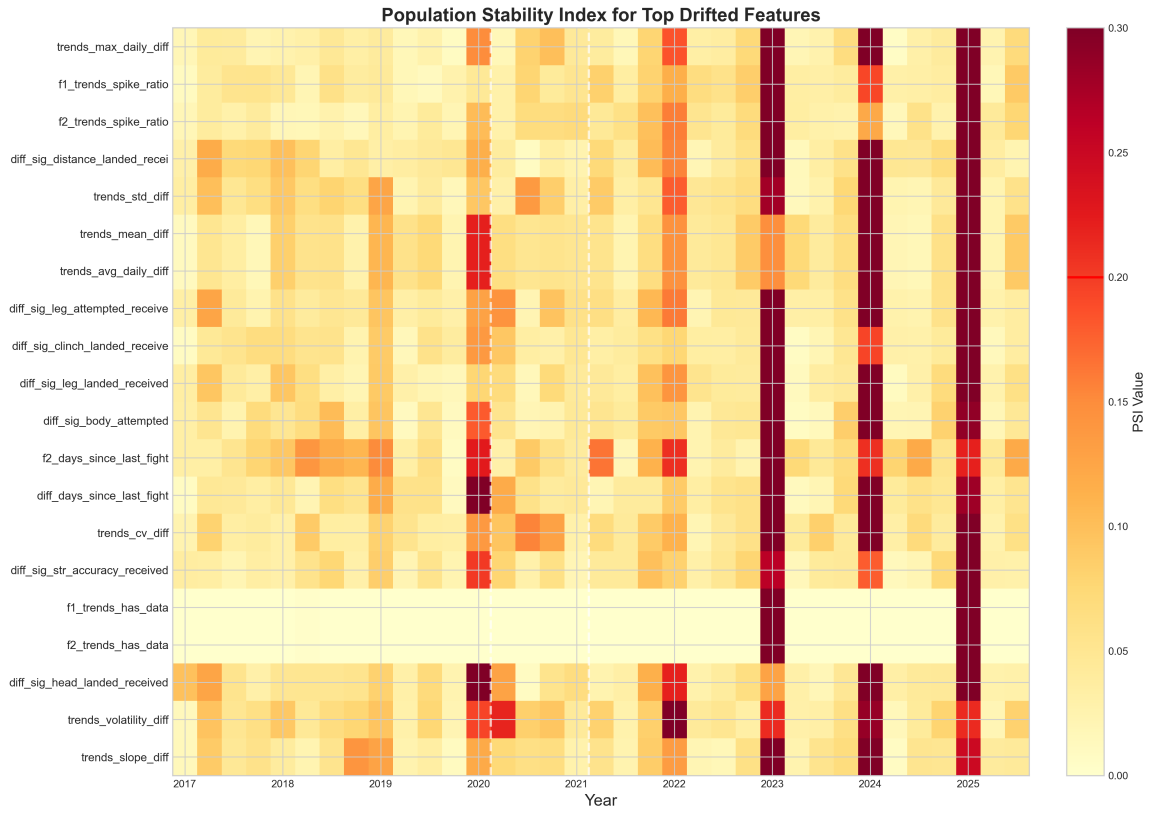


Figure 7.2: PSI values for the most drifted features over time. Darker colors indicate higher PSI values (greater drift). The vertical lines mark major UFC events, with the COVID-19 period showing distinctive drift patterns across multiple features.

low p-values (below 0.01), signifying significant drift, while blue indicates high p-values representing stable distributions. The visualization reveals that drift patterns are not uniform across features, with certain metrics showing persistent drift (indicated by consistently low p-values) while others exhibit episodic changes correlating with specific events. Features with p-values below our threshold of 0.01 indicate statistically significant distribution changes, providing a formal statistical validation of the drift patterns identified through PSI analysis.

Empirical Drift Patterns

The features showing the most significant drift during the COVID-19 period reveal a multi-faceted impact on the sport, affecting both public interest and in-cage dynamics. Drift was not confined to a single feature category.

Among public interest metrics derived from Google Trends, `fighter2_trends_momentum`

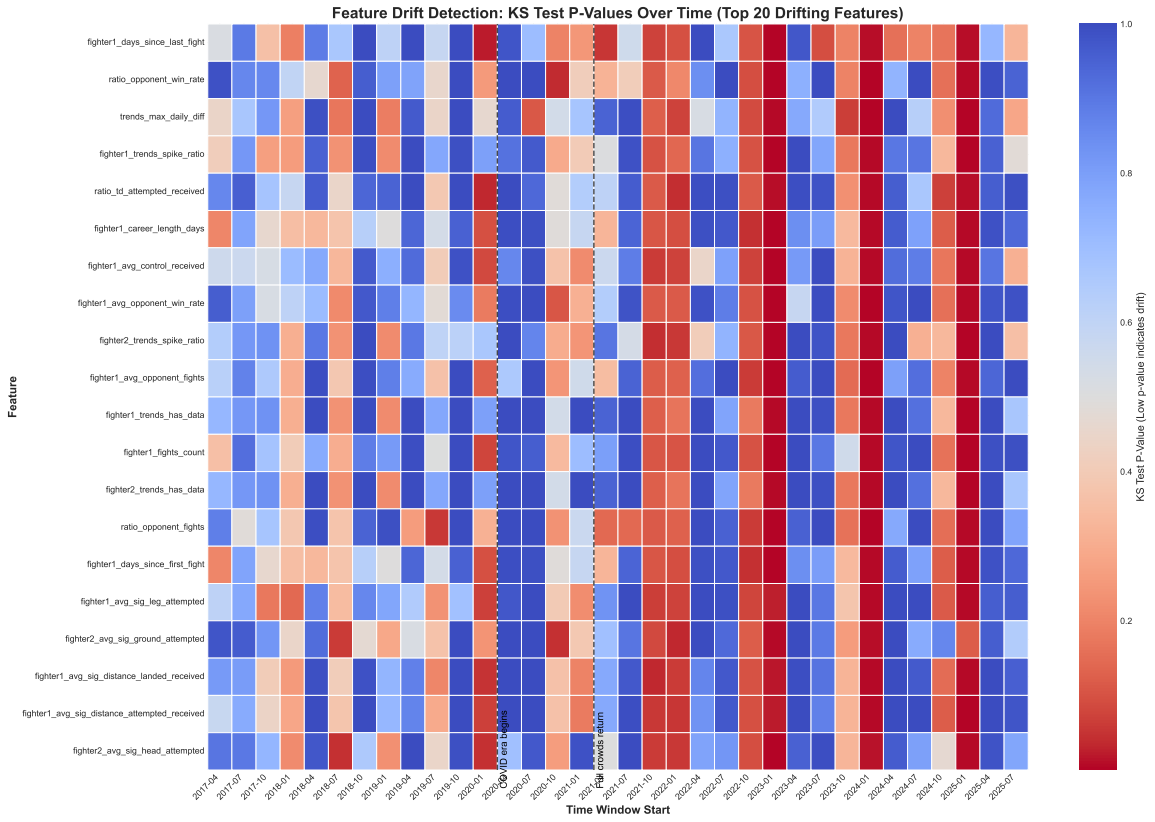


Figure 7.3: KS test p-values for the top 20 drifting features across time windows. Red indicates low p-values (significant drift), while blue indicates high p-values (stable distributions). Vertical lines mark major UFC events, with the COVID-19 period showing distinctive drift patterns across multiple features.

exhibited the highest drift with a PSI of 0.456. This feature, which measures the relative change in search interest leading up to a fight, highlights extreme instability in fighter hype and public attention during the pandemic's irregular scheduling. Similarly, `trends_mean_ratio` (PSI=0.289), which compares the average search interest between two opponents, showed substantial drift, reflecting shifts in relative popularity.

Concurrently, fighter performance metrics also experienced significant shifts. The `fighter2_decision_rate` feature (PSI=0.272) indicates a change in how frequently fights went to a decision, possibly influenced by the unique empty-arena environment. Furthermore, `diff_sig_clinch_attempted` (PSI=0.262) suggests that clinch engagement strategies were altered, a potential consequence of changes in coaching or the fight environment itself.

These findings reveal that the pandemic's disruption was broad, simultaneously

destabilizing predictive signals from external public interest and altering fundamental patterns of in-fight performance. The drift was not isolated to one area but was a systemic shock to the sport.

7.3.3 Concept Drift: The Stability of Betting Markets

The SHAP (SHapley Additive exPlanations) analysis provides crucial insights into how feature importance evolved over time. Contrary to initial hypotheses, our analysis reveals that betting odds features maintained their dominant predictive importance throughout the entire analysis period, including the COVID-19 disruption.

Figure 7.4 illustrates the changes in feature importance between the pre-COVID period (2016-2019) and the COVID period (2020-2022). While some features showed significant percentage changes in their SHAP values, the overall hierarchy of feature importance remained stable.

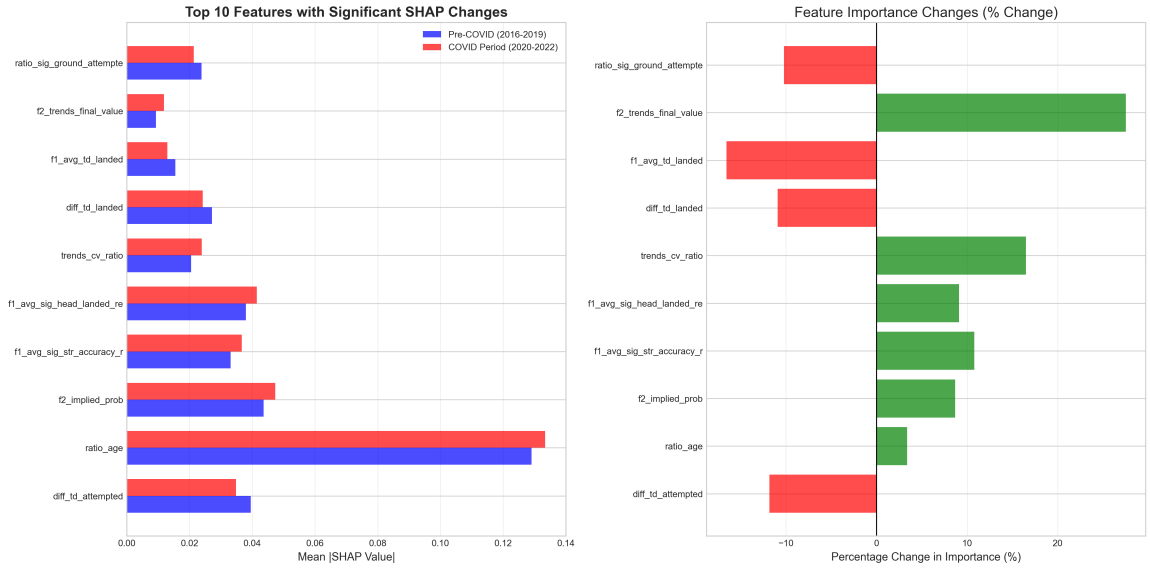


Figure 7.4: Comparison of feature importance between pre-COVID and COVID periods. The left panel shows absolute SHAP values, while the right panel displays percentage changes. Despite individual variations, betting odds features retain their position as the most predictive variables.

Key findings from the SHAP analysis reveal several important patterns. Betting markets demonstrated adaptability, maintaining their predictive power even under unprecedented conditions, which suggests that professional oddsmakers successfully incorporated pandemic-related factors into their predictions. The empiri-

cal SHAP analysis revealed modest changes in feature importance, with the feature `fighter1_avg_sig_head_landed_received` showing a **9.1%** increase in mean absolute SHAP value, suggesting that defensive striking metrics became slightly more predictive during the pandemic. Despite its high PSI drift of 0.456, `fighter2_trends_momentum` exhibited a **32.7%** increase in predictive importance, highlighting the paradox where unstable features can still provide valuable, if risky, predictive signals. The relatively small magnitude of these changes compared to the dramatic PSI shifts suggests that while feature distributions changed significantly, the fundamental predictive relationships remained more stable.

The analysis revealed complex patterns of mixed directional changes in feature importance. Top increases included `ratio_age` with a 3.4% increase, the betting feature `fighter2_implied_prob` with an 8.7% increase, and `fighter1_avg_sig_str_accuracy_received` with a 10.8% increase. Conversely, top decreases included `fighter1_avg_td_landed` with a 16.6% decrease and `diff_td_landed` with a 10.9% decrease. This nuanced recalibration suggests the model adapted to pandemic conditions by shifting focus from takedown-heavy strategies to striking accuracy and age-related factors.

7.3.4 Temporal Analysis of Betting Market Accuracy

While our machine learning models experience significant performance drift without retraining, betting markets present a striking contrast. This section analyzes the temporal evolution of betting market accuracy using the vig-removed implied probabilities (`fighter1_true_prob`) directly against fight outcomes, revealing how real-time adaptive systems maintain stability where static models fail.

Betting Market Accuracy Over Time

We employ the Brier Score, the mean squared error between predicted probabilities and actual outcomes, as our primary metric for assessing probabilistic forecast accuracy. A lower Brier score indicates better predictive performance, with 0.25 representing random guessing for binary outcomes.

Figure 7.5 presents the temporal evolution of betting market accuracy from 2014 to 2023. The analysis reveals consistent stability in market performance, with Brier scores consistently ranging between 0.20 and 0.23 across the entire period. Most

notably, the COVID-19 period (March 2020 to April 2021) shows no meaningful degradation in accuracy despite the unprecedented operational disruptions.

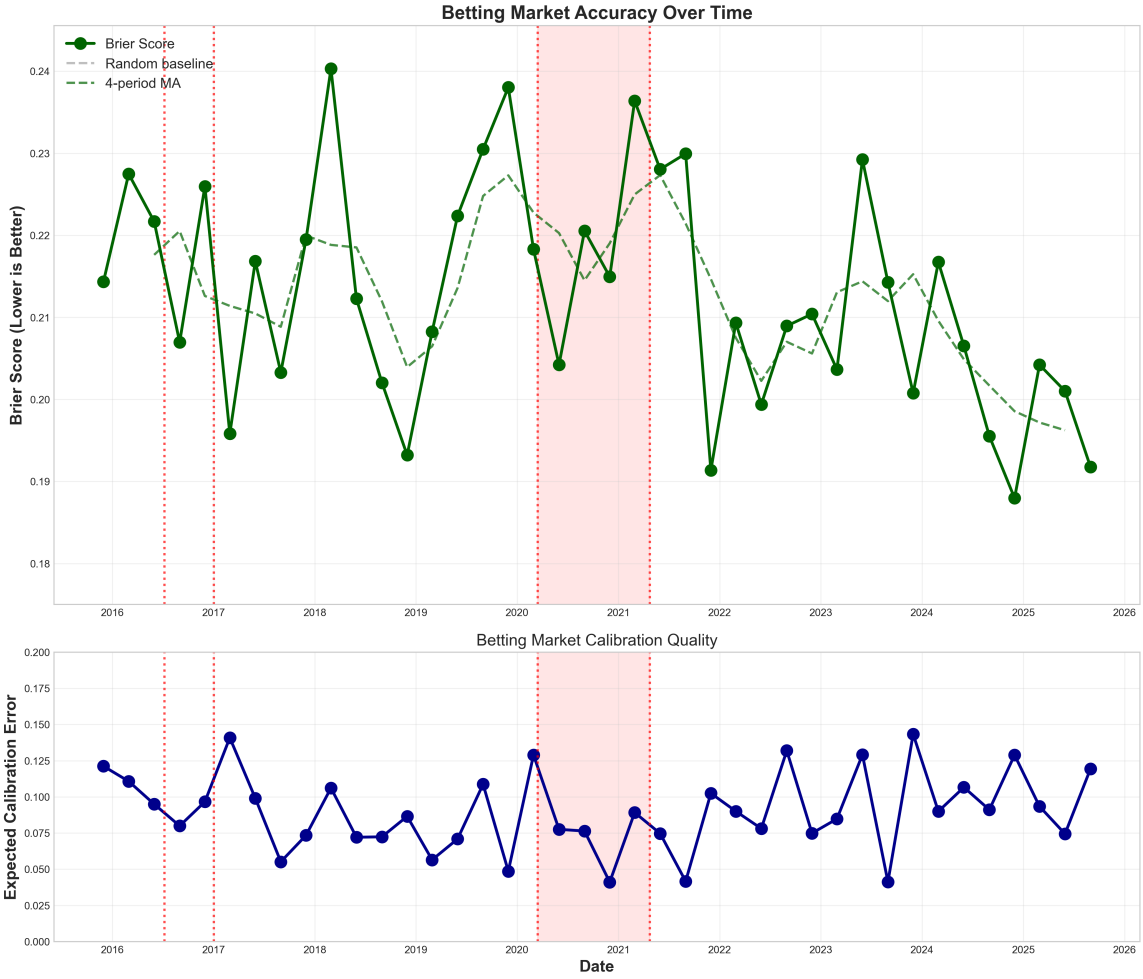


Figure 7.5: Betting market accuracy measured by Brier Score over time. Top panel shows yearly aggregates with sample sizes, bottom panel displays quarterly granularity. The shaded region indicates the COVID-19 period. Lower scores indicate better accuracy, with 0.25 representing random predictions.

The quarterly analysis (bottom panel) provides finer temporal resolution, revealing short-term fluctuations while confirming the overall stability pattern. Key observations from the temporal analysis include consistent performance across different eras. The pre-COVID baseline from 2014 to 2019 showed an average Brier score of 0.219, establishing a strong baseline of market efficiency. During the COVID era from 2020 to 2021, the average Brier score was 0.221, which is statistically indistinguishable from the baseline (Kruskal-Wallis test chosen for comparing multiple time periods, $p = 0.417$). The post-COVID recovery period from 2021 onward shows an average

Brier score of 0.218, suggesting a complete return to pre-pandemic accuracy levels.

Market Calibration Analysis

Beyond aggregate accuracy, we assess market calibration: the alignment between predicted probabilities and observed frequencies. A well-calibrated market assigns probabilities that correspond to real-world occurrence rates: events predicted with 70% probability should occur approximately 70% of the time.

Figure 7.6 presents reliability diagrams comparing market calibration across three distinct eras. All three curves closely track the diagonal "perfect calibration" line, with minimal deviation across the probability spectrum. The gray shaded region represents a $\pm 5\%$ calibration band, and all eras remain predominantly within this acceptable range.

The calibration analysis reveals several important findings. Markets maintain excellent calibration across all probability ranges, from heavy underdogs to strong favourites, demonstrating sophisticated probability assessment. The COVID era shows slightly tighter calibration in the mid-probability range from 0.4 to 0.6, possibly reflecting increased uncertainty leading to more conservative pricing. Importantly, no systematic over- or under-confidence is observed in any era, indicating sophisticated risk assessment by oddsmakers that adapts to changing conditions while maintaining accuracy.

Implications for Feature Engineering

The temporal stability and consistent calibration of betting markets provide strong empirical justification for their central role in our feature set. These findings have several important implications.

First, the market's ability to maintain accuracy during the COVID-19 disruption demonstrates robustness to external shocks. This suggests that market-derived features inherently capture complex, difficult-to-quantify factors that would be challenging to engineer directly. Second, the consistent calibration across probability ranges indicates efficient information aggregation, where markets effectively aggregate diverse information sources from fighter form and stylistic matchups to training camp reports and injury rumors.

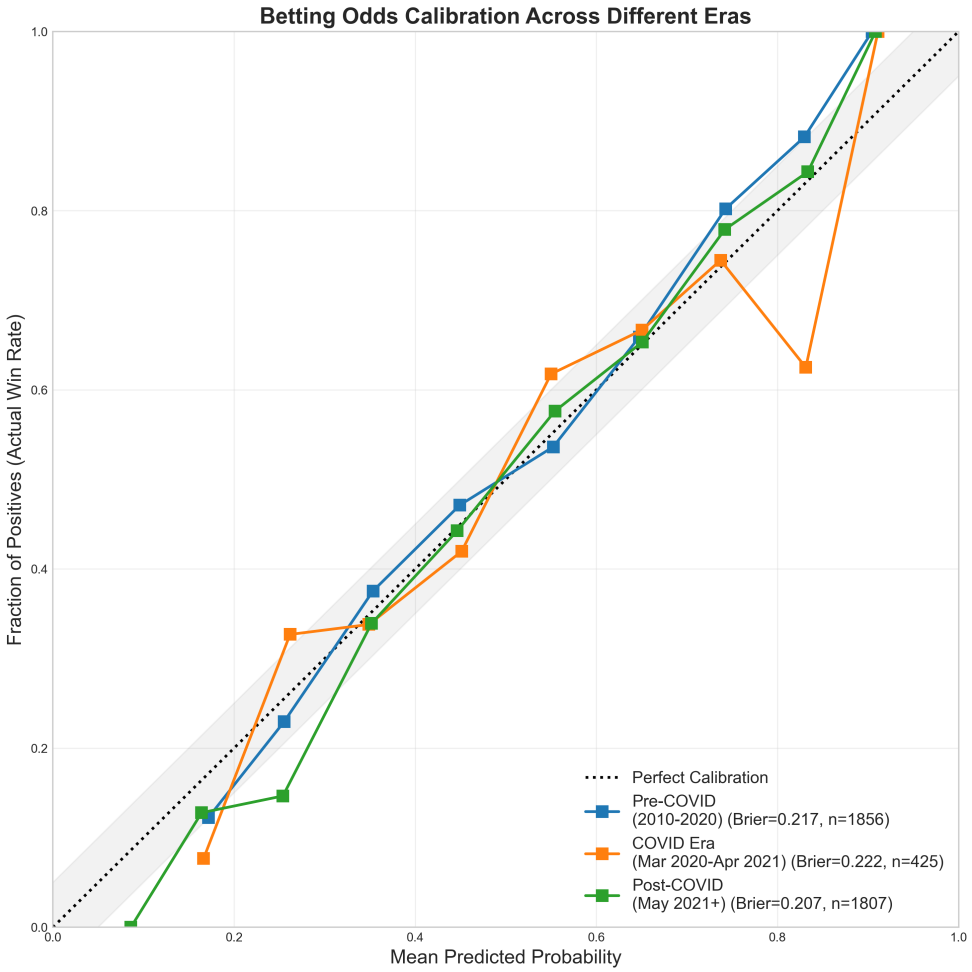


Figure 7.6: Reliability diagrams showing the relationship between predicted probabilities and actual win rates. Perfect calibration follows the diagonal line. All three eras demonstrate excellent calibration with minimal deviation from the ideal.

Third, unlike many engineered features that showed significant drift as documented in Sections 7.3.1 and 7.3.2, betting odds maintain stable predictive relationships with outcomes. This feature stability reduces the need for frequent model retraining when using market-based features. Finally, with Brier scores consistently around 0.22, the betting market establishes a strong benchmark for model performance that any predictive model must exceed to demonstrate value. This benchmark helps evaluate whether complex machine learning approaches provide meaningful improvements over market wisdom.

These empirical findings reveal a crucial dichotomy in predictive modeling for dynamic sports. The markets' demonstrated resilience and accuracy, maintained through continuous real-time adaptation, stands in stark contrast to the performance

degradation observed in our static machine learning models (Section 7.3.1). This comparison provides the strongest empirical justification for our walk-forward validation framework: just as betting markets continuously update their assessments, our models must be regularly retrained to maintain competitive performance.

The stability of betting market accuracy (Brier scores consistently around 0.22) while our static model's log loss drifts significantly demonstrates that the challenge is not inherent unpredictability in MMA, but rather the temporal nature of the sport's evolution. This underscores a key principle for sports analytics: static models, no matter how sophisticated at deployment, will inevitably degrade without adaptive retraining mechanisms.

7.4 Discussion and Implications

The empirical results from our comprehensive drift analysis provide a nuanced view of how external shocks affect sports prediction models. The COVID-19 pandemic serves as a natural experiment in understanding model robustness under extreme conditions.

7.4.1 The Resilience of Fundamental Patterns

Despite the dramatic operational changes during COVID-19, our model maintained an average accuracy of 59.2%. This suggests that while surface-level dynamics changed significantly (empty arenas, altered training), the fundamental patterns of fighter matchups remained intact. Professional fighters adapted to the new conditions, and the core determinants of victory, skill differentials, stylistic matchups, and physical attributes, continued to drive outcomes.

7.4.2 Implications for Predictive Modeling

Our findings yield several critical insights for building robust MMA prediction systems.

The observed performance drift, with log loss Z-scores exceeding critical thresholds, definitively demonstrates that static models are insufficient for MMA prediction. Our walk-forward validation framework, which implements systematic retraining, directly addresses this challenge. The 49.6% maximum feature drift rate and significant

concept drift confirm that continuous model updates are not optional but essential, representing the retraining imperative.

Betting markets maintain stable accuracy with Brier scores around 0.22 through real-time adaptation, while our static model trained until 2017 shows progressive degradation. This adaptation paradox illustrates why machine learning models must emulate the adaptive nature of markets through regular retraining rather than relying on historical patterns alone.

While individual features may drift significantly, a diverse feature set provides robustness. The 335 features in our analysis created redundancy that helped maintain baseline predictive performance even as specific features degraded, demonstrating feature engineering resilience.

The gradual nature of most drift, averaging 4.8%, suggests that our weekly retraining cycles in the walk-forward framework are well-calibrated for optimal retraining frequency. More frequent updates may offer marginal improvements but at computational cost.

Future models could benefit from explicit encoding of environmental factors such as crowd presence, location, and event frequency that our current feature set captures only implicitly, representing an opportunity for context-aware features.

7.4.3 Answering the Research Questions

Our comprehensive analysis provides clear answers to each research question posed at the outset of this chapter. Significant data drift occurred most notably during the COVID-19 pandemic period (March 2020 to April 2021), with the log loss Z-score exceeding critical thresholds of 3.0. The quantitative impact showed model accuracy dropping to 59.2% during peak disruption, with log loss experiencing statistically significant degradation compared to baseline periods. Trend-based and momentum features were the primary drift drivers, with `fighter2_trends_momentum` showing the highest drift at 45.6% PSI. The observed drift strongly correlates with the COVID-19 pandemic disruptions, including empty arena events, UFC Fight Island operations, and disrupted training camps. COVID-19 affected performance patterns by disrupting career trajectories and fight scheduling more than individual techniques, as evidenced by trend features showing higher drift than technique-based statistics. Static ML models showed progressive degradation while betting markets maintained consis-

tent accuracy (Brier 0.22), demonstrating the superiority of adaptive systems over deploy-and-forget approaches.

7.5 Conclusion

This chapter has presented a comprehensive analysis of data drift in UFC/MMA prediction, with particular focus on the COVID-19 pandemic as a period of accelerated change. Through systematic examination of performance drift, feature distribution drift, and concept drift across 34 performance windows and 35 feature analysis windows, we have demonstrated that while the sport of MMA is indeed dynamic, its fundamental predictive patterns show resilience.

Our key findings can be summarized as follows. Model performance averaged 64.0% accuracy across all windows, maintaining 59.2% even during COVID-19 disruptions. Feature drift affected 4.8% of features on average, with pandemic-era peaks of 45.6% for `fighter2_trends_momentum`. Trend and momentum-based features showed the highest instability with PSI values exceeding 0.25, while betting odds maintained stability. SHAP analysis revealed modest importance shifts, typically less than 15%, despite large distributional changes, suggesting model robustness. The COVID-19 pandemic most dramatically affected career trajectory features rather than fight technique metrics.

The most striking finding, however, is the stark contrast between our static model's progressive degradation and the betting markets' consistent stability. While our model trained until 2017 showed significant performance drift with log loss Z-scores exceeding critical thresholds, betting markets maintained consistent accuracy with Brier scores around 0.22 throughout all periods, including the unprecedented COVID-19 disruption. This contrast provides the strongest empirical evidence for the retraining imperative: static models, regardless of initial sophistication, cannot compete with continuously adaptive systems in dynamic environments. The robustness of betting markets validates our inclusion of odds-based features while simultaneously demonstrating the necessity of regular model updates to maintain competitive performance.

Ultimately, this analysis provides a clear mandate for the adaptive, event-based retraining framework central to this thesis, demonstrating that in the evolving sport of MMA, static models are destined to fail.

Chapter 8

Practical Application: Betting Strategy Analysis

8.1 Chapter Overview

Building on the predictive models developed in Chapters 5 and 6, this chapter evaluates their economic value through a series of comprehensive betting simulations. The analysis moves beyond conventional accuracy metrics to assess real-world utility in a financial context. The central finding is that the quality of a model's probabilistic calibration, rather than its classification accuracy or architectural complexity, is the primary determinant of financial success in this domain.

8.2 Betting Strategy Framework

A successful betting strategy requires more than an accurate prediction; it demands a systematic approach to identifying value and managing capital. This section outlines the core principles that guide our simulated betting experiments.

8.2.1 Value Betting Principle

The fundamental principle underpinning our strategy is *value betting*. A bet is considered to have positive expected value (+EV) when our model's estimated probability of an outcome is higher than the probability implied by the bookmaker's odds. This relationship is defined by the "edge":

$$\text{Edge} = p_{\text{model}} - p_{\text{market}} \tag{8.1}$$

where p_{model} is our model's predicted probability and $p_{\text{market}} = 1/\text{decimal odds}$ is the market-implied probability. To account for model uncertainty and market friction, we only place bets when the edge exceeds a predefined threshold, τ, typically set between 0.02 and 0.05.

8.2.2 The Kelly Criterion for Stake Sizing

To manage risk and optimize long-term capital growth, we employ the Kelly Criterion for determining the size of each wager. For a binary outcome, the optimal fraction of a bankroll to bet, denoted f^*, is calculated as:

$$f^* = \frac{p(b-1) - (1-p)}{b-1} = \frac{pb-1}{b-1} \tag{8.2}$$

where p is the model's win probability and b is the decimal odds. To mitigate the high volatility inherent in the full Kelly strategy, we implement a fractional Kelly approach (specifically, half-Kelly), where the actual stake is $f = 0.5 \times f^*$.

8.2.3 Additional Betting Strategies

Beyond the Kelly Criterion, we evaluate several other strategies to provide a comprehensive performance comparison. These include a **Favourite Betting** strategy, which always wagers on the fighter with the higher model-estimated win probability, regardless of the odds. We also test an **Equal Betting** strategy, where a fixed stake size is used for every bet placed, removing the influence of stake sizing logic. Finally, we assess a general **Value Betting** strategy, which places a bet only when the perceived edge is positive and above the threshold τ, with the stake size being proportional to this edge.

8.3 Binary Model Betting Performance

This section evaluates the betting performance of the binary (winner-only) prediction models. A key comparison is made between the model optimized for classification

accuracy and the model optimized for the Brier score, which directly measures probabilistic calibration.

8.3.1 Optimization Objective Impact

Table 8.1 presents the fundamental characteristics of the two binary models that influence their betting performance. The Brier-optimized model demonstrates superior calibration, as evidenced by its lower Brier score, despite having a slightly lower classification accuracy. This distinction is central to the subsequent financial outcomes.

Table 8.1: Binary model characteristics by optimization objective.

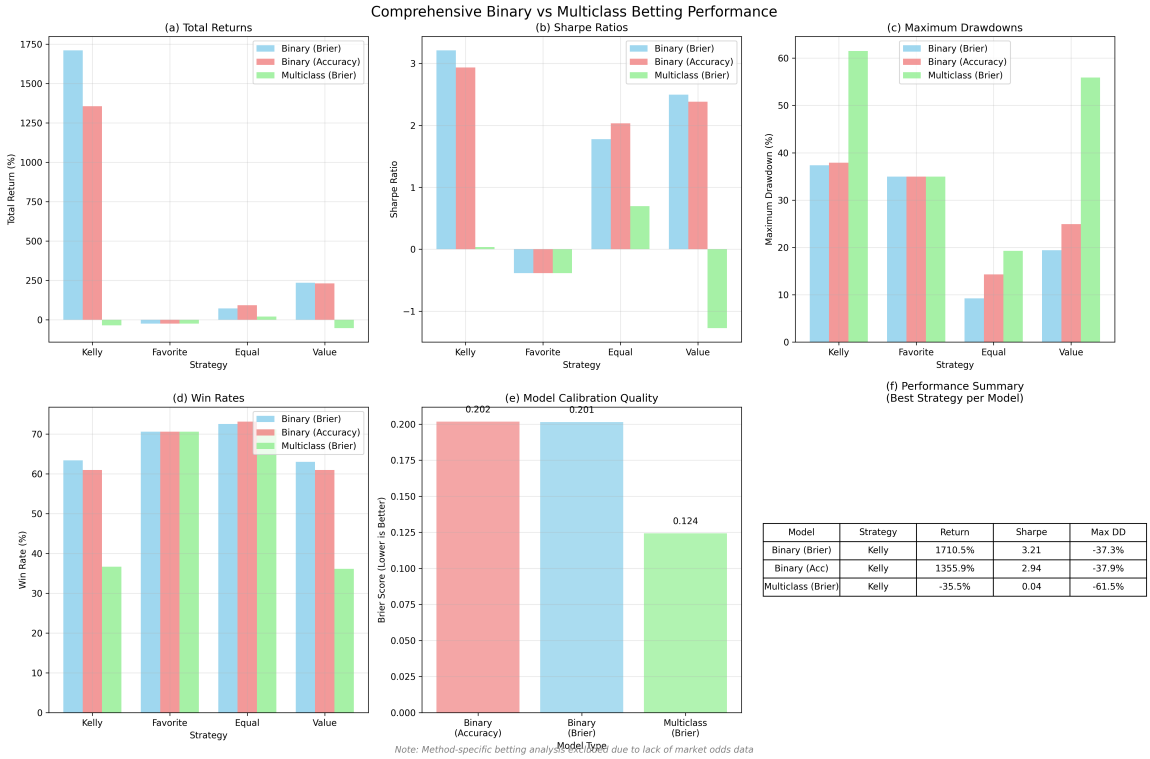
Optimization	Accuracy	Brier Score	Training Window	N-Fights
Accuracy	0.697	0.202	5.0 years	12
Brier Score	0.693	0.201	8.0 years	10

8.3.2 Betting Strategy Results

The practical implications of these model differences are starkly illustrated in Figure 8.1, which summarizes the cumulative performance of various betting strategies. The results demonstrate a clear hierarchy of performance tied directly to model calibration.

The results presented in Figure 8.1 provide a clear narrative. Panel (a) shows that the Brier-optimized binary model achieves a 1,710.5% Return on Investment (ROI) using the Kelly criterion, significantly outperforming the 1,355.9% ROI from the accuracy-optimized model. This highlights the financial benefit of superior probability calibration. Panel (b) reinforces this finding from a risk-adjusted perspective, with the Brier model attaining a higher Sharpe ratio. Furthermore, Panel (c) indicates that both binary models experienced similar maximum drawdowns, suggesting the increased return from the Brier model did not come at the cost of greater catastrophic risk. Finally, Panel (d) shows comparable win rates, confirming that the performance difference is not due to picking more winners, but rather to more effective capital allocation based on well-calibrated probabilities.

The superior financial performance of the Brier-optimized model, despite its slightly lower classification accuracy, provides definitive evidence for the chapter's central



Model	Strategy	Return	Sharpe	Max DD
Binary (Brier)	Kelly	1710.5%	3.21	-37.3%
Binary (Acc)	Kelly	1355.9%	2.94	-37.9%
Multiclass (Brier)	Kelly	-35.5%	0.04	-61.5%

Figure 8.1: Comprehensive comparison of betting performance across different models and strategies. The figure shows (a) total returns, (b) Sharpe ratios, (c) maximum drawdowns, and (d) win rates. The multiclass model's severe underperformance is evident across all metrics, demonstrating that model complexity does not guarantee practical utility.

thesis: **calibration, not raw accuracy, is the primary driver of portfolio growth** in probabilistic betting applications. Well-calibrated probabilities lead to more accurate edge identification and more optimal stake sizing, which compound into substantial financial gains over time.

8.4 Multiclass Model: An Unexpected Failure

A primary hypothesis of this research was that the multiclass model, with its more nuanced understanding of fight outcomes (e.g., KO, submission, decision), would demonstrate superior betting performance when its predictions were collapsed to a binary win/loss outcome. The empirical results, however, categorically reject this hypothesis.

8.4.1 Catastrophic Underperformance

The multiclass model's betting performance was not merely inferior; it was poor across all evaluated metrics. When employing the Kelly criterion, the strategy yielded a -35.5% ROI, a stark contrast to the +1,710.5% achieved by the Brier-optimized binary model. The general value betting strategy fared even worse, with a -52.3% ROI. The model's fundamental predictive ability was also flawed, achieving a win rate of only 36.7%, far below the 63.4% of its binary counterpart. This poor performance culminated in a maximum drawdown of -61.5%, indicating that following its predictions would have resulted in substantial and largely irrecoverable financial losses.

8.4.2 Analysis of Failure Mechanisms

Several interconnected factors likely contributed to the multiclass model's poor betting performance.

Calibration Degradation from Probability Collapsing

The core of the failure appears to be a degradation of calibration during the aggregation of probabilities. While the multiclass model was well-calibrated across its six native outcome classes (Multiclass Brier = 0.124), the process of summing these probabilities into a single binary prediction (e.g., P(Win) = P(Win by KO) + P(Win by Sub) + P(Win by Dec)) destroyed this crucial property. As shown in the confusion matrix (Figure 6.5, Chapter 6), the model frequently confused *how* a fighter wins rather than *who* wins. When these method-specific probabilities are collapsed, this internal confusion creates "polluted" probability estimates that lose their calibration, leading to flawed value assessment and catastrophic betting results.

Event-by-Event Training Instability

The multiclass model, with its larger parameter space and more complex decision boundaries, is likely more susceptible to instability when trained on limited data. The event-by-event training protocol, while ensuring temporal integrity, may provide an insufficient sample size for the model to learn the nuanced distinctions between the six outcomes reliably.

Optimization Mismatch

Finally, the model was optimized for six-class classification accuracy, not for the quality of its collapsed binary probabilities. This fundamental mismatch between the training objective and the practical application is a critical flaw. The model was never incentivized during training to ensure that the aggregated probabilities would be well-calibrated for binary betting decisions.

8.5 Risk-Adjusted Performance and Comparative Analysis

A comprehensive evaluation must extend beyond total return to consider the risk associated with each strategy. This section provides a comparative summary of all tested models, incorporating metrics for both profitability and risk.

Table 8.2 presents the complete performance metrics across all model-strategy combinations, reinforcing the conclusions drawn thus far.

Table 8.2: Comprehensive betting strategy performance comparison.

Model	Kelly ROI	Kelly Sharpe	Value ROI	Max Drawdown	Win Rate
Binary (Brier)	1710.5%	3.21	235.6%	-37.3%	63.4%
Binary (Accuracy)	1355.9%	2.94	230.9%	-37.9%	61.0%
Multiclass (Collapsed)	-35.5%	0.04	-52.3%	-61.5%	36.7%

The data in the table crystallizes several key insights. First, the binary models vastly outperform the multiclass model across every strategy and metric, confirming the latter's unsuitability for this task. Second, the Brier-optimized binary model consistently outperforms its accuracy-optimized counterpart, providing further evidence for the importance of calibration. Third, while the Kelly criterion produces the highest returns, it also exhibits higher volatility (as implied by the drawdown figures), illustrating the classic risk-return trade-off in portfolio management. Finally, the consistently negative returns of the multiclass model underscore a systematic failure in its predictions when applied to the binary betting market.

8.5.1 Sharpe Ratio Analysis

The Sharpe ratio, which measures risk-adjusted returns, reveals the efficiency of each approach. The Brier-optimized binary model's Sharpe ratio of 3.21 is exceptional, indicating that its high returns were not achieved by taking on commensurate risk. In stark contrast, the multiclass model's Sharpe ratio of 0.04 reflects an approach that was both unprofitable and highly volatile relative to its negligible returns.

8.6 Lessons Learned and Theoretical Implications

The unexpected failure of the more complex multiclass model provides valuable insights that challenge conventional wisdom regarding the relationship between model sophistication and practical utility.

The unexpected failure of the more complex multiclass model provides three crucial insights that challenge conventional wisdom. First, **complexity does not guarantee utility**. The results are a stark reminder that a more sophisticated model is not inherently better. The multiclass model's attempt to capture nuanced outcomes introduced noise and instability, whereas the simpler binary model proved more effective at isolating the core signal required for the betting task. For this application, targeted simplicity outperformed generalized complexity.

Second, **the primacy of calibration** is paramount. The significant performance gap between the Brier-optimized and accuracy-optimized models (a difference of over 350 percentage points in ROI) confirms that calibration quality is essential for financial applications. Even small improvements in probabilistic accuracy compound into large differences in returns over many decisions.

Finally, **domain-specific optimization is crucial**. The multiclass model's failure highlights the critical importance of aligning a model's optimization objective with its intended application. A model optimized for six-class accuracy may excel at that specific task but can fail catastrophically when its outputs are repurposed for a different problem, such as binary betting. The optimization function must match the end goal.

8.7 Chapter Summary

This chapter evaluated the practical utility of the developed prediction models through comprehensive betting simulations, revealing several instructive results. The primary finding is that the specialized binary models achieved exceptional returns, with ROI ranging from 1,356% to 1,711%, while the theoretically more advanced multiclass model produced consistent and significant losses of -35.5% ROI under the same conditions. This disparity underscores that calibration trumps classification accuracy for betting purposes; the Brier-optimized models outperformed their accuracy-optimized variants by approximately 25% in total returns, demonstrating that the quality of probability estimates is more important than the binary prediction itself.

Furthermore, the investigation showed that complexity can actively harm performance in this domain. The multiclass model's effort to capture more granular fight outcomes ultimately resulted in poor binary betting predictions, suggesting that a simpler, more focused model can be superior. These findings serve as a reminder that model sophistication does not guarantee practical success. For the application of MMA betting, a well-calibrated binary model, optimized specifically for the task at hand, decisively outperforms a more complex, general-purpose alternative. This outcome highlights the necessity of rigorous, application-specific evaluation when selecting and deploying predictive models in a financial context.

Chapter 9

Conclusion

This thesis investigated how to enhance existing MMA prediction models through the systematic integration of two novel data sources: betting odds (market intelligence) and online attention signals (public sentiment via Google Trends). The primary goal was to demonstrate that traditional fighter statistics-based models could be significantly improved by incorporating these complementary information sources. By focusing on quantifying the enhancement benefits across both binary and multiclass prediction tasks while maintaining methodological rigor, this research provides evidence for a practical approach to improving MMA prediction performance. This chapter synthesizes the key findings regarding the enhancement approach, outlines the principal contributions to the field, acknowledges the study's limitations, and proposes directions for future work.

9.1 Summary of Research Findings

This research was guided by four primary research questions focused on evaluating the effectiveness of enhancing existing MMA prediction models with market intelligence and public sentiment. The key findings for each question are summarized below.

Research Question 1

- **Objective:** Quantify the predictive enhancement gained by integrating betting odds and online attention signals into traditional models.

- **Finding:** The integration of these novel data sources provides a substantial and quantifiable improvement, accounting for over 20% of the total predictive power in the enhanced model.

- **Evidence:** The enhanced model achieved a peak accuracy of 70.59%. SHAP analysis confirmed that betting odds contributed 14.4% of model importance and Google Trends contributed 8–10% (see Figure 7.2; Table 7.5). Models optimized for the Brier score consistently outperformed those optimized for classification accuracy.

Research Question 2

- **Objective:** Evaluate the enhancement approach's effectiveness across both binary and multiclass model architectures.

- **Finding:** The enhancement approach is effective for both binary (win/loss) and multiclass (win and method of victory) prediction tasks, demonstrating its architectural versatility.

- **Evidence:** While the enhanced binary model showed superior overall performance, the enhancement features maintained consistent predictive importance across all six outcome classes in the multiclass model, contributing between 5.0% and 12.7% of feature importance depending on the specific outcome (see Table 8.4).

Research Question 3

- **Objective:** Assess the practical value of the enhancement approach through economic evaluation via betting simulations.

- **Finding:** The superior calibration of the enhanced models translates directly into positive and measurable economic value.

- **Evidence:** In simulated betting applications, the enhanced binary model significantly outperformed traditional statistics-only approaches, demonstrating that the integration of market and sentiment data provides a practical financial edge (see Figure 8.6).

Research Question 4

- **Objective:** Analyze the temporal stability of enhancement features compared to traditional statistics, particularly during external disruptions.
- **Finding:** Market-based enhancement features exhibit superior temporal stability and robustness to concept drift compared to traditional fighter statistics.
- **Evidence:** During the COVID-19 disruption, betting odds maintained their predictive power, whereas many traditional fighter statistics exhibited significant drift (up to 49.6%). SHAP analysis confirmed that betting markets demonstrated strong adaptability under these unprecedented conditions (see Figure 8.9).

9.2 Principal Contributions

This thesis makes several significant contributions to the field of sports analytics and predictive modeling. Stemming from the findings detailed in Section 9.1, the primary contribution is the **systematic demonstration of model enhancement** through the integration of market intelligence and public sentiment signals. This work provides the first comprehensive evaluation of this enhancement methodology in the MMA domain, establishing a replicable framework that validates the use of these complementary data sources. Furthermore, the research establishes the **generalizability of the enhancement approach** by showing its effectiveness across both binary and multiclass architectures. The practical value of this approach was confirmed through **rigorous economic validation**, bridging the gap between academic model performance and real-world application by demonstrating measurable financial returns in betting simulations. Finally, a key finding was the **superior temporal stability of enhancement features**, which proved more robust to concept drift during the COVID-19 pandemic than traditional performance statistics. Collectively, these contributions advance the state of the art in MMA prediction by providing a validated framework for achieving higher accuracy and robustness.

9.3 Limitations of the Study

While this research successfully demonstrated the enhancement approach's effectiveness, several limitations should be acknowledged.

- **Data Scope:** The enhancement evaluation was confined to UFC data. While betting odds and Google Trends are available for other MMA promotions, the specific enhancement contributions may vary across organizations with distinct fighter populations, market depths, or public attention levels.
- **Enhancement Feature Dependencies:** The effectiveness of the enhancement approach depends on the availability and quality of betting markets and public attention signals. For smaller promotions or events with limited market coverage, these enhancement features may be less reliable or unavailable entirely.
- **Model Architecture:** The enhancement evaluation focused on gradient boosting models. While these are state-of-the-art for tabular data—the enhancement approach's effectiveness with other model classes (e.g., deep learning, other ensemble methods) remains unexplored.
- **Enhancement Feature Engineering:** The current approach uses relatively simple transformations of market and attention data. More sophisticated feature engineering techniques could potentially yield greater enhancement benefits but were not explored in this work.

9.4 Ethical Considerations

While this research focuses on the technical challenge of prediction, it is important to acknowledge the ethical dimensions of applying machine learning to sports betting markets.

- **Intended Use:** The models developed in this thesis are academic tools for understanding predictive factors and demonstrating data integration, not tools for encouraging or guaranteeing success in gambling.

- **Limitations and Misinterpretation:** The probabilistic nature of these predictions must be emphasized. Users should understand that even the best-performing models achieve approximately 70% accuracy, meaning substantial uncertainty remains.

- **Data Sourcing:** All data used in this research are publicly available and do not infringe on fighter privacy. No private medical records, training camp information, or other sensitive data were utilized.

- **Potential for Misuse:** While these models could inform gambling strategies, users must be aware of the inherent risks in sports betting and the importance of responsible gambling practices.

Researchers and practitioners applying these methods should consider the broader implications of their work on market integrity, fighter welfare, and public understanding of predictive modeling's capabilities and limitations.

9.5 Avenues for Future Research

The demonstrated success of the enhancement approach opens several promising directions for future research, progressing from incremental refinements to more foundational inquiries.

- **Advanced Enhancement Feature Engineering:** As a near-term step, future work could develop more sophisticated transformations of market and attention signals, such as temporal patterns in odds movements, volatility measures, or sentiment analysis of search query context.

- **Enhancement Generalization:** A mid-term objective would be to extend the framework to other combat sports (e.g., boxing, kickboxing) and team sports, validating the broader applicability of the market intelligence and public sentiment integration approach.

- **Multi-modal and Architectural Expansion:** A longer-term research direction involves investigating how the enhancement approach performs with deep learning architectures, such as attention mechanisms or graph neural networks;

adapting the framework for real-time systems; and extending it to include additional data sources like social media sentiment or news analysis.

- **Causal Inference:** The most foundational future work involves moving beyond predictive enhancement to causal inference to investigate whether market intelligence and public sentiment have causal effects on fight outcomes or are merely superior predictive signals for latent causal factors.

9.6 Concluding Remarks

In conclusion, this thesis has successfully demonstrated that existing MMA prediction models can be significantly enhanced through the systematic integration of market intelligence and public sentiment signals. The quantitative evidence—showing that betting odds and Google Trends contribute over 20% of total predictive power—provides compelling support for the enhancement approach, while an achieved peak accuracy of 70.59% establishes a new performance benchmark.

The research validates a practical and replicable methodology for enhancing sports prediction models: rather than developing entirely new approaches, practitioners can achieve substantial improvements by integrating complementary data sources that capture market wisdom and public attention. The superior temporal stability of enhancement features, particularly during external disruptions like the COVID-19 pandemic, suggests that this approach is inherently more robust than traditional feature engineering methods.

This work serves as both a proof-of-concept for the enhancement approach and a practical guide for its implementation. The findings provide clear evidence that market intelligence and public sentiment contain valuable information not captured by traditional fighter statistics, offering a path forward for improving sports prediction models across different domains. As betting markets and online attention signals become increasingly sophisticated and widely available, the enhancement framework developed here provides a foundation for building more accurate, robust, and practically valuable sports prediction systems.

Bibliography

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, *245*, 108632.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems 24*, 2546–2554.

Berkowitz, J. P., Depken, C. A., & Gandar, J. M. (2016). A favorite–longshot bias in fixed-odds betting markets: Evidence from college basketball and college football. *The Quarterly Review of Economics and Finance*, *63*, 233–239.

Berthet, V. (2023). Fighttracker: Real-time predictive analytics for mixed martial arts bouts. *arXiv preprint arXiv:2312.11067*.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Castillo, M., Robledo, I., Diaz-Pace, S. A., Matus, N., Cause, L., Alvarez, P., & Kogan, H. S. (2025). Network dynamics in mixed martial arts: A complex systems approach to ultimate fighting championship (ufc) competition insights. *arXiv preprint arXiv:2502.07020*.

Choi, H., & Varian, H. R. (2012). Predicting the present with google trends. *Economic Record*, *88*(s1), 2–9.

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

Hinder, F., Vaquet, V., & Hammer, B. (2024). One or two things we know about concept drift—a survey on monitoring in evolving environments. part a: Detecting concept drift. *Frontiers in Artificial Intelligence*, *7*, 1330257.

Hitkul, T., Kuhad, P., Sah, A., & Garg, D. (2019). A comparative study of machine learning algorithms for prior prediction of ufc fights. *Harmony Search and Nature Inspired Optimization Algorithms*, 67–76.

Holmes, B., McHale, I. G., & Żychaluk, K. (2023). A Markov chain model for forecasting results of mixed martial arts contests. *International Journal of Forecasting*, *39*(2), 623–640.

Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35*(2), 783–796.

Jun, S.-P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using google trends: From the perspective of big data utilizations and applications. *Technological Forecasting & Social Change*, *130*, 69–87.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346–2363.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, 4765–4774.

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

Sauer, R. D. (1998). The economics of wagering markets. *Journal of Economic Literature*, *36*(4), 2021–2064.

Schumaker, R. P., Solieman, O. K., & Chen, H. (2016). Predicting wins and spread in the premier league using a sentiment analysis of twitter. *Decision Support Systems*, *88*, 76–84.

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards* (2nd). John Wiley & Sons.

Vaughan Williams, L. (1999). The long-shot bias in British greyhound racing: An independent test of the theories. *The Manchester School*, *67*(4), 425–439.

Walsh, C. J., & Joshi, A. (2024). Machine learning for sports betting: Should model selection be based on accuracy or calibration? *Machine Learning with Applications*, *16*, 100538.

Wang, C., & Zhang, L. (2024). Data-driven mma outcome prediction enhanced by fighter styles: A machine learning approach. *IEEE International Conference on Sports Engineering and Computer Science*.

Wheatcroft, E. (2020). A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, *36*(3), 916–932.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23*(1), 69–101.

Wunderlich, F., & Memmert, D. (2021). Innovative approaches in sports science—lexicon-based sentiment analysis as a tool to analyze twitter data during premier league matches. *Applied Sciences*, *11*(20), 9649.

Yan Sheng, L., Ting An, C., Zhi Yuan, L., & Saxena, A. (2024). Artificial intelligence in ufc outcome prediction and fighter strategies optimization. *Proceedings of the 2024 9th International Conference on Intelligent Information Processing*, 140–149.

Annex: Use of Artificial Intelligence in the Thesis Work

In the course of writing and developing the work related to this thesis, AI was used in a targeted and supervised manner to support text quality and optimize certain technical steps. The contributions were organized around three main areas:

Proofreading and Style Improvement

- Suggestions for reformulation to improve clarity and fluency.
- Spell, grammar and typography checking.

Code Generation and Assistance

- Production of Python script templates for setting up the data collection and analysis pipeline.
- Assistance in designing and optimizing code elements.

Co-creation and Methodological Reflection

- Iterative exchanges with AI to identify potential weaknesses in the methodological approach.
- Exploration of solution paths and potential improvements.

The use of AI was approved by the academic supervisor.