Leveraging Alternative Data for Mixed Martial Arts Betting Markets

Eugene Han

January 22, 2025

Advised by: Brian Macdonald and Robert Wooster, Statistics & Data Science

Introduction

Sports betting has experienced a dramatic rise in popularity and financial volume over the past decade, fueled by increasing legalization, the proliferation of online platforms, and widespread media coverage. In 2022 alone, the global sports betting market was valued at over $83 billion, with projections estimating continued growth at an annual rate of 10.3% through 2030.

Among sports gaining traction globally, mixed martial arts (MMA)—particularly through the Ultimate Fighting Championship (UFC), the sport's leading promotion—has seen exponential growth in both viewership and fan engagement. This surge in popularity has naturally extended to betting markets, where wagers on UFC fights now constitute a significant share of betting activity.

Despite its popularity, modeling UFC fights presents unique challenges. Unlike team sports with well-established datasets and long histories of play, MMA data is both scarce and sparse. Fighter performance metrics are often incomplete or inconsistent, and external variables such as weight cuts, injuries, and stylistic mismatches contribute to the sport's inherent volatility. These factors make accurately predicting outcomes particularly challenging and suggest the presence of inefficiencies in the betting markets. This project explores the potential to capitalize on these inefficiencies using novel approaches with respect to both data and methodology.

Problem Statement

Problem

Most public research and existing projects related to UFC betting markets focus exclusively on data from a single source, UFC Stats, due to its ease of access and comprehensive fight statistics. While valuable, this narrow focus may overlook additional insights that could be gained from incorporating a broader range of data sources. My approach seeks to address this limitation by leveraging diverse and unconventional datasets alongside UFC Stats to uncover new signals and potential inefficiencies in the betting markets. Furthermore, I aim to experiment with innovative modeling techniques and betting strategies, including methods inspired by conformal prediction and robust optimization.

Data

The following ten (10) data sources will be considered, loosely organized into groups:

- **Striking/Grappling Metrics** - UFC Stats, ESPN
- **Betting Odds** - Best Fight Odds, FightOdds.io
- **Complete Fighting Histories** - Sherdog, Fight Matrix
- **Judge Scoring** - MMA Decisions
- **Miscellaneous** - Wikipedia, Bet MMA, Tapology

As of January 15, 2025, all needed data from these sources have been scraped or obtained from someone else's work (such as an existing GitHub repository). Although most of the data cleaning was already handled by the scraping pipelines implemented, some of the files require additional attention to address edge cases.

In addition to cleaning the data, the following two data-related tasks need to be completed:

1. *Cross-Source Matching*. Almost all of the listed data sources have their own unique set of IDs assigned to events, fights, and fighters. All of these must be matched to each other through a set of central "link" tables so that unified train and test datasets can be easily created. This will most likely involve a mix of (fuzzy) string matching, metadata matching, and some manual labor.
2. *Exploratory Data Analysis and Feature Engineering*. Even after cleaning and matching, the data cannot be used directly for modeling since there is a mix of information known before any given fight and information that is known only after a fight has taken place. Features must be created from useful information that informs the outcome of a fight and relies only on data before the corresponding fight takes place. While most of these features will be created based on intuition/domain knowledge with EDA not being the main focus of this project, some degree of analysis and visualization will be needed to sanity check ideas.

Methods

The core idea is to estimate the win probabilities of each fighter in a given fight based only on information that would be available at inference time and to bet on outcomes where those predicted probabilities deviate sufficiently from those implied by the betting odds.

We will accomplish this by framing the problem as a binary classification task and experimenting with logistic regression and tree-based models, as well as the inclusion of calibration layers using Platt scaling, isotonic regression, and Venn-Abers predictors. In particular, Venn-Abers predictors will output a tuple of probabilities that serve as a sort of confidence interval with validity guarantees that reflect the uncertainty in a prediction.

The bet sizes will be computed using an extension of the fractional Kelly criterion to simultaneous outcomes, experimenting with different fractions such as 0.5, 0.25, and 0.1. For pipelines using Venn-Abers calibration, we will also use a robust version of the Kelly criterion that considers the uncertainty in the probability estimates to optimize for the worst-case scenario.

To validate the presence (or lack thereof) of an edge, we will perform backtests over fights within the last several years and examine different frequencies of model retraining.

Deliverables

- Final database(s) with cleaned data and "link" tables with IDs matched across sources
- Universal train and test datasets to be used by all modeling pipelines
- Trained models
- Backtesting results for all method/approach combinations mentioned in Methods
- Poster and report

Timeline

- Week 1: Proposal, set up meeting schedule with advisors
- Weeks 2-5: Finalize meeting schedule with advisors, start writing (draft abstract, introduction, data sections), clean datasets, match IDs across sources, load data to SQLite databases
- Weeks 6-8: EDA and feature engineering, continue writing (discuss EDA and feature engineering process), start implementing model training pipelines
- Spring Break: Wrap up feature engineering, finish implementing model pipelines
- Weeks 9-10: Train all models, implement backtesting framework, continue writing (modeling, backtesting)
- Weeks 11-13: Run backtests, wrap up drafts of report and poster, revise based on feedback
- Reading Week: Finalize report and submit