Applying Machine Learning and Convex Optimization to the UFC Betting Market

Eugene Han
Department of Statistics
Yale University
e.han@yale.edu

1 Problem Statement

Mixed martial arts has seen a rapid increase in popularity in mainstream sports, with the Ultimate Fighting Championship (UFC) being one of the largest promotions in the world. Consequently, betting on UFC events has become increasingly prominent, intensified by aggressive marketing and easy-to-use platforms provided by bookmakers like DraftKings and FanDuel.

Betting markets are generally very efficient and the implied probabilities of sportsbooks' odds are often “accurate" in the sense that they match up well with the relative frequency of events occurring. However, bookmakers can offer mispriced odds under/overestimating a fighter's chances of winning, yielding opportunities to generate profit if identified correctly.

This project will be a preliminary exploration into whether or not one can obtain an edge in the UFC betting market by producing well-calibrated probabilities for fight outcomes and formulating an optimization problem to calculate optimal wager sizes using those predicted probabilities.

Past work related to this has largely focused on using rather generic and easily accessible data from the UFC Stats website; we will consider more niche and unconventional datasets from third-party websites, a novel approach relative to public research on forecasting UFC fights.

2 Approach

Our approach will consist of the following components:

1. *Alternative Data* - We will be web scraping a wide variety of data from several disparate sources, some of which are niche and contain information that can't be found anywhere else. See section 3 for details on each source.
2. *Feature Engineering* - Creating a rich feature set will likely be the most time-consuming part of this project as most of the scraped data is essentially unusable without some transformation and aggregation. Care will be taken to prevent data leakage, ensuring all features are constructed using information known strictly before a fight occurs.
3. *Modeling* - We will construct an L2-penalized logistic regression model with the regularization parameter chosen through time series cross-validation, trained on fights between 2010 and 2022. Since this is a preliminary attempt at the problem, a good baseline would be logistic regression; moreover, logistic regression tends to be well-calibrated by default. An L2 penalty is chosen due to a potentially large set of candidate features.
4. *Bet Sizing* - We extend the Kelly criterion for optimal bet sizing to a general number of simultaneous bets. Given there are n fights in an upcoming fight and restricting to moneyline straight bets (no parlays, spreads, or over/under bets), there exist $2n + 1$ potential positions–two outcomes for each fight and a risk-free asset (no bet). Maximizing for expected log

wealth, we can formulate this as the following optimization problem:

$$\begin{array}{ll} \max\limits_{b} & \pi^T \log\left(Rb\right) \\ \text{s.t.} & 1^T b = 1 \\ & b \succeq 0 \end{array}$$

where $\pi \in \mathbb{R}^{2^n}$ is a vector of probabilities corresponding to each of the 2^n different event outcomes, $R \in \mathbb{R}^{2^n \times (2n+1)}$ is the returns matrix encoding the returns for each of the $2n+1$ positions for each of the 2^n outcomes, and $b \in \mathbb{R}^{2n+1}$ is a vector of wager proportions for each bet. This problem is always feasible: in the worst case, take $b^* = e_{2n+1}$, the $(2n+1)$-th basis vector in $\mathbb{R}^{2n+1}$, which is equivalent to not betting at all and guarantees a profit/loss of 0. Since π is intractable, we can compute an estimate $\hat{\pi}$ using the predicted probabilities from our classifier (given it is well-calibrated) and assuming independence of fights within an event.

5. *Backtesting* - Finally, we will evaluate our system's betting performance on unseen fights in 2023 and 2024 given average closing odds, which reflect the most information and are therefore the "sharpest" lines. Caution will be taken to prevent leakage and all fights will be considered, including those that ended in draws or no contests, to simulate bets that may be voided and thus preserve realism.

3 Data

This project will leverage data from the following sources:

- UFC Stats - Main analytics website for the UFC that contains per-round fight statistics regarding aspects like striking and takedowns as well as fighter characteristics such as height and date of birth
- Sherdog - Third-party website not affiliated with the UFC that acts as a database for 300,000+ fights across multiple professional and amateur promotions; has information on UFC fighters' full professional careers including fights before signing with the UFC and extra data like nationalities
- Fight Matrix - Third-party website that uses Sherdog's data and proprietary algorithms to maintain a custom ranking system and Elo-like ratings
- FightOdds.io - Historical and upcoming betting odds across several bookmakers
- Open-Elevation API - Open-source API for getting elevation data rounded to the nearest meter for a given set of latitude and longitude coordinates

All the raw data has been scraped through another one of the author's projects using the Scrapy framework.

This is a good way of demonstrating my approach since we will be considering many different kinds of data, offering room to be creative in the feature engineering process. Data scarcity will not be a huge issue, since the training set will contain several thousand fights. Moreover, the historical closing odds will allow us to get realistic, conservative estimates of betting performance.

4 Success Metrics

Success will be determined in terms of how the model compares to closing odds as well as the profitability of the entire system. Since closing lines theoretically capture the most information known before a fight, we would ideally like to achieve a log loss as close as possible to or even better than the bookmakers' implied probabilities–no worse than by 0.05 could be a starting point. For profitability, success would be a positive return over the backtesting period.

5 Milestones and Deliverables

Major milestones for the project are as follows:

1. Create candidate features/sets of features for each data source
2. Train and save logistic regression model
3. Evaluate log loss and betting performance on fights in 2023 and 2024

Deliverables will include a complete dataset, a trained and saved logistic regression model, a performance comparison between the model and the closing odds for the backtest period, and an evaluation of betting returns on that same period.