Leveraging Alternative Data for Mixed Martial Arts Betting Markets

Eugene Han

Yale University, Department of Statistics & Data Science

Introduction

Beating the betting market for UFC fights is challenging due to outcome volatility and sparse data. Prior studies rely solely on the UFC Stats website, suffer from small backtesting samples, and have methodological flaws. Our work overcomes these limitations by curating a novel dataset from nontraditional sources and performing rigorous backtests, while also exploring new modeling and betting strategies inspired by robust optimization and conformal prediction.

Dataset Curation

Data was scraped from 10 websites, cleaned, and standardized across sources, resulting in a 411 MB relational database with 58 tables, 6.9 million rows, and 64.7 million individual data points, $> 50\times$ larger than UFC Stats alone (8 MB).

- **Striking/Grappling**: UFC Stats, ESPN
- **Betting Odds**: Best Fight Odds, FightOdds.io
- **Fight History/Rankings/Ratings**: Fight Matrix, Sherdog
- **Judge Scoring**: MMA Decisions
- **Miscellaneous**: Bet MMA, Tapology, Wikipedia

Feature Engineering

- **Event-Level**: Shared attributes across same-event fights (e.g., venue elevation)
- **Bout-Level**: Attributes of individual fights (e.g., weight class)
- **Fighter-Level**: Comparative measures based on fighters' attributes and past performance (e.g., difference in historical strikes landed per second)

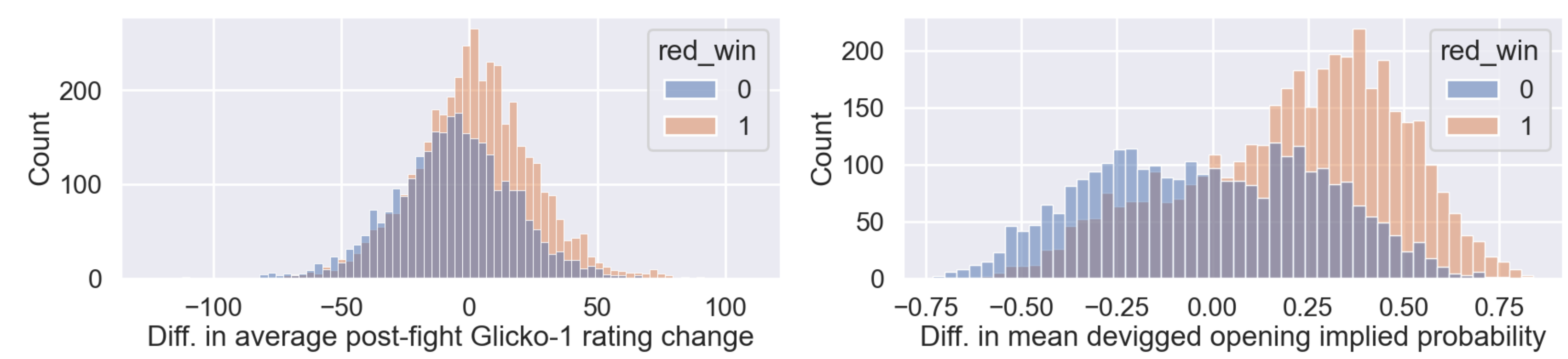

Figure 1. Class distributions for two examples of predictive features

Modeling Approach

Let $Y_i = \mathbb{1}\{\text{red corner fighter wins fight } i\}$. **Goal**: Model $\mathbb{P}(Y_i = 1 \mid X_i)$.

1. **Odds Feature Ablation**: Inclusion/exclusion of opening odds-derived feature
2. **Feature Selection**: Variance thresholding followed by top K selection via mutual information scores
3. **Base Model**: Experimented with ridge logistic regression and gradient boosting
4. **Calibration**: Inclusion/exclusion of using Venn-Abers predictors, which outputs an interval (p_0, p_1) with validity guarantees such that $p_0 \leq \mathbb{P}(y = 1 \mid x) \leq p_1$

Betting Strategies

Suppose an event has m fights. There exist $2m+1$ bets and 2^m possible outcomes.

Simultaneous Kelly: Find allocation b, to maximize expected log growth rate of wealth, $G_\pi(b) = \pi^T \log(R^T b)$. For $m = 2$, construct R and estimate π as

$$R = \begin{pmatrix} o_{r,1} & o_{r,1} & 0 & 0 \\ 0 & 0 & o_{b,1} & o_{b,1} \\ o_{r,2} & 0 & o_{r,2} & 0 \\ 0 & o_{b,2} & 0 & o_{b,2} \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad \hat{\pi} = \begin{pmatrix} \hat{p}_1 \hat{p}_2 \\ \hat{p}_1(1-\hat{p}_2) \\ (1-\hat{p}_1)\hat{p}_2 \\ (1-\hat{p}_1)(1-\hat{p}_2) \end{pmatrix}$$

given odds $o_r = (o_{r,1}, \ldots, o_{r,m})$, $o_b = (o_{b,1}, \ldots, o_{b,m})$ and $\hat{p}_j = \hat{\mathbb{P}}(Y_j = 1 \mid X_j)$.

Distributional Robust Kelly: Maximize expected worst case log growth rate, $G_\Pi(b) = \inf_{\pi \in \Pi} G_\pi(b)$, over uncertainty set $\Pi = \{\pi \in \Delta_{2^m} \mid A\pi \leq d\}$ with

$$A = \begin{pmatrix} -I_{2^m} \\ I_{2^m} \end{pmatrix}, \quad d = \begin{pmatrix} -\hat{\pi}_{\text{lower}} \\ \hat{\pi}_{\text{upper}} \end{pmatrix}$$

where $\hat{\pi}_{\text{lower}}, \hat{\pi}_{\text{upper}}$ are defined like $\hat{\pi}$ using the (p_0, p_1) outputs from Venn-Abers.

Backtesting Setup

- **Date Range**: 8-year period, 2017 to 2024 (3960 bouts, 331 events)
- **Training/Tuning**: Refit after every event, retune at end of each year
- **Bankroll Details**: Initial = \$1000, Kelly fraction $f \in \{0.10, 0.15, 0.25\}$
- **Benchmark**: Closing odds from Bovada Sportsbook
- **Significance Testing**: Monte Carlo simulations using closing odds with

$$\text{p-value} = \frac{(\text{\# of simulations with profit} \geq \text{observed}) + 1}{(\text{total \# of simulations}) + 1}$$

and adjusted using a Bonferroni correction

Model Results

Model Pipeline	Log Loss	Brier Score
Logistic Regression	**0.608060**	0.210443
Logistic Regression (No Odds)	0.629998	0.220371
Venn-Abers Logistic Regression	0.608881	0.210849
Venn-Abers Logistic Regression (No Odds)	0.632596	0.221477
LightGBM	0.611128	0.211772
LightGBM (No Odds)	0.631593	0.221174
Venn-Abers LightGBM	0.611205	0.211956
Venn-Abers LightGBM (No Odds)	0.633977	0.222121
Bovada Sportsbook	*0.608142*	***0.210265***

Table 1. Summary of model metrics over the backtest period, compared with closing odds

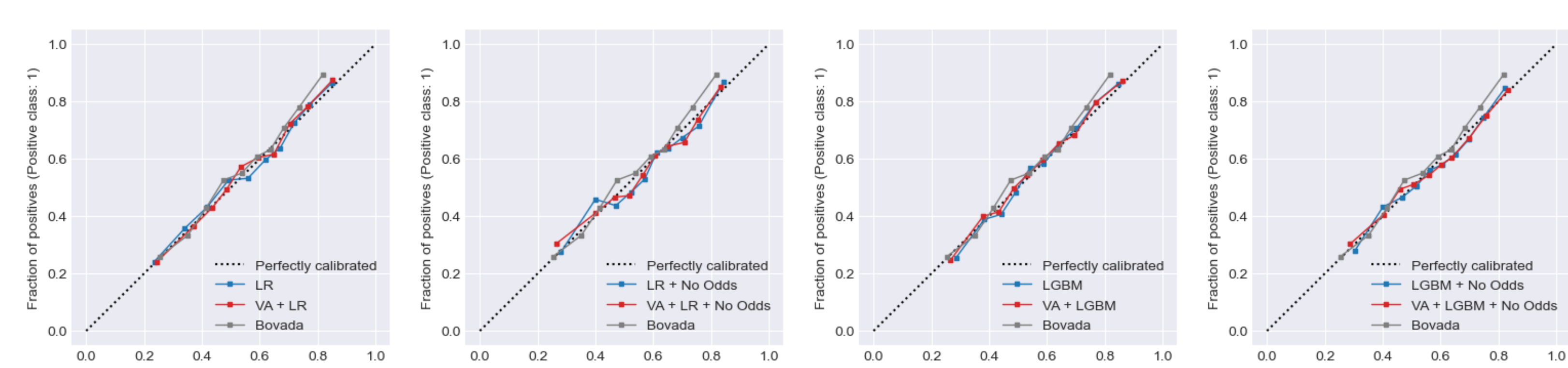

Figure 2. Calibration plots comparing model pipelines with and without Venn-Abers

Model Results (cont.)

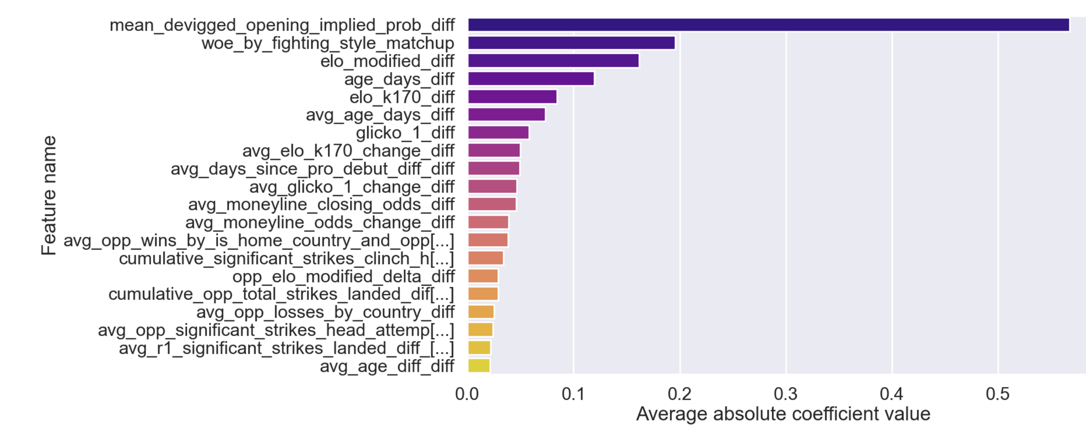

Figure 3. Top 20 features by average absolute coefficient value in logistic regression model

Betting Results

Model Pipeline	Betting Strategy	Fraction	Profit (\$)	Total Bets	Yield (%)	MDD (%)	Adj. p-value
Logistic Regression	Simultaneous	0.10	1318.98	1178	6.40	-22.04	**0.0516**
		0.15	2159.40	1196	5.25	-32.02	
		0.25	**3763.63**	1201	3.47	-50.46	
Logistic Regression (No Odds)	Simultaneous	0.10	-114.95	1810	-0.53	-40.19	0.9083
Venn-Abers Logistic Regression	Simultaneous	0.10	430.80	1445	2.24	-34.74	0.3528
	Distributional Robust	0.10	338.04	609	**7.47**	**-16.36**	0.3780
Venn-Abers Logistic Regression (No Odds)	Simultaneous	0.10	-181.54	1865	-0.78	-45.45	1.0000
	Distributional Robust	0.10	-260.40	1396	-2.28	-34.03	1.0000
LightGBM	Simultaneous	0.10	192.37	1610	0.89	-49.20	0.5999
LightGBM (No Odds)	Simultaneous	0.10	-584.14	1906	-3.29	-68.99	1.0000
Venn-Abers LightGBM	Simultaneous	0.10	258.96	1540	1.21	-46.78	0.5111
	Distributional Robust	0.10	130.44	881	1.93	-30.03	1.0000
Venn-Abers LightGBM (No Odds)	Simultaneous	0.10	-551.64	1894	-2.86	-72.54	1.0000
	Distributional Robust	0.10	-394.33	1380	-2.57	-62.51	1.0000

Table 2. Betting metrics by model pipeline and betting strategy combination

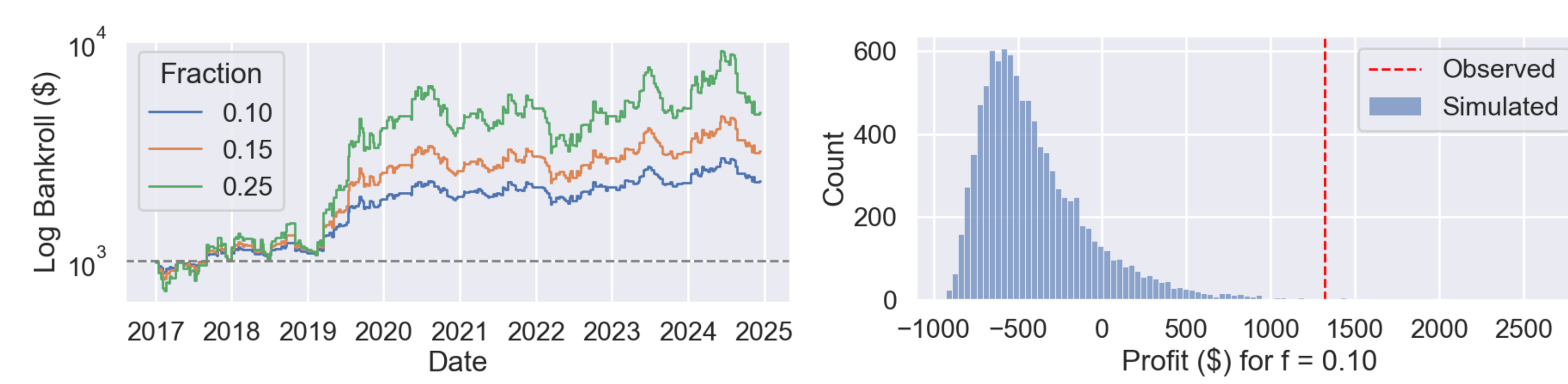

Figure 4. Backtest performance for logistic regression and simultaneous Kelly

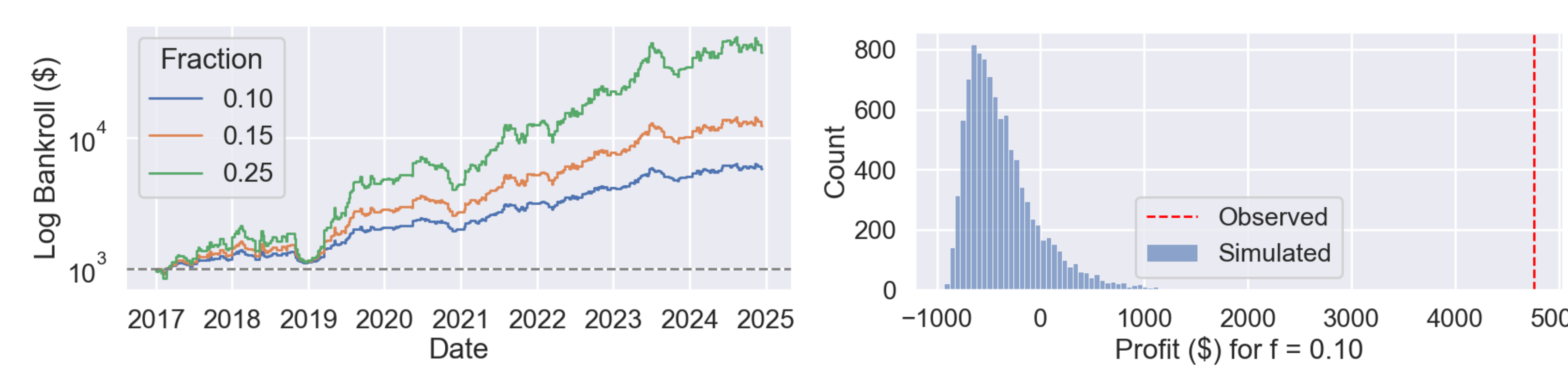

Figure 5. Backtest performance with women's and debut fights removed

References

[1] Mikoláš Bartoš. Machine learning in combat sports. Bachelor's thesis, Czech Technical University in Prague, 2021.

[2] Qingyun Sun and Stephen Boyd. Distributional Robust Kelly Gambling: Optimal Strategy under Uncertainty in the Long-Run, 2021.

[3] Vladimir Vovk and Ivan Petej. Venn-abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 829–838, Arlington, Virginia, USA, 2014. AUAI Press.

Github: ehan03

CSAS 2025 Poster Session

Email: e.han@yale.edu