Leveraging Alternative Data for Mixed Martial Arts Betting Markets

Yale Sports Analytics

Eugene Han

Department of Statistics & Data Science
Yale University

April 18, 2025

Mixed Martial Arts (MMA)

- "Hybrid combat sport incorporating techniques from boxing, wrestling, judo, jujitsu, karate, Muay Thai (Thai boxing), and other disciplines"
- Ultimate Fighting Championship (UFC) is the largest global promotion $\rightarrow$ focus of this project

Sports Betting

- Moneyline bets: wagers placed on a game's outcome, i.e. who will win
  - Voided if there is no winner
- Odds are a dual representation
  - Payouts
  - Implied probabilities
- Example: Fighter A ($+150$) vs. Fighter B (-170)
  - \$2.50 returned for every \$1 wagered if fighter A wins, implies 40% win chance
  - \$1.59 returned for every \$1 wagered if fighter B wins, implies 63% win chance
  - House edge of 3%

Main Takeaway

Beating the market is a **probabilistic** problem!

Problem Definition

For a given fight i, let Y_i be defined as

$$Y_i = \begin{cases} 1 & \text{if the red corner fighter wins} \\ 0 & \text{if the blue corner fighter wins} \end{cases}$$

We want to

1. Model $\mathbb{P}(Y_i = 1 \mid X_i)$ using information known strictly before the event
2. Place bets when our model's predicted probabilities disagree with those implied by the betting market

Previous Works

- Some existing work, mostly amateur projects
- Common limitations:
  - Almost exclusively uses UFC Stats, the official data provider of the UFC
  - Methodology issues with modeling and/or feature engineering
  - No attention to calibration (e.g., do model's probabilities actually reflect confidence/empirical proportions?)
  - Unrealistic backtesting setups, too small of sample sizes
- How can we improve on this?

Innovations

1. Novel dataset incorporating "alternative" (i.e. nontraditional) data sources
2. Rigorous backtesting over 8 years with hypothesis testing
3. Experimenting with ideas from conformal prediction framework and robust optimization for betting under uncertainty

The Dataset

Source	Examples of Data
UFC Stats	Per-round striking/grappling stats
ESPN	Additional bout-aggregated striking/grappling stats
Wikipedia	Event attendance, venue location/capacity
Sherdog	Fighters' full professional fight histories
Fight Matrix	Custom monthly rankings and Elo-like rating scores
Tapology	Weigh-in results, gym/team affiliations
MMA Decisions	Per-round scoring by judge
Best Fight Odds	Historical timestamped betting odds for older events
FightOdds.io	Historical betting odds for recent years, fighting styles
Bet MMA	Weight misses, short notice fights

Table: Overview of data sources

The Dataset (cont.)

- About 2 to 2.5 months of work
  - Scraping, cleaning, cross-source ID matching
- Relational database design (SQL $>$ CSVs/data frames)
  - 58 tables
  - 6.9 million rows
  - 64.7 million individual data points
  - 411 MB, over 50$\times$ bigger than UFC Stats alone ($\sim$ 8 MB)
- Completely open-source

Feature Engineering

- "Kitchen sink" approach: Generate tons of candidates, throw away most later
  - Mostly based on domain knowledge/intuition
  - Semi-systematic
- Three "groupings"
  - **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)
- **Important**: All features were created respecting temporal order

Feature Engineering (cont.)

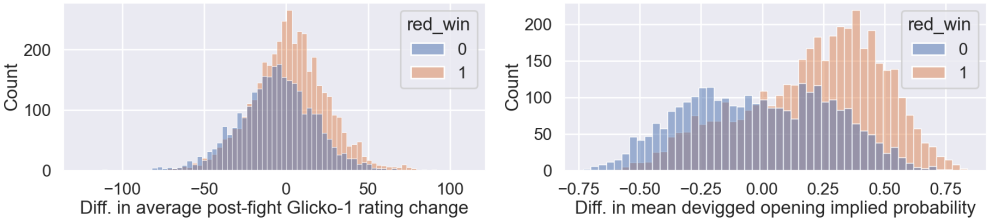


Figure: Class distributions for example fighter-level features

Modeling

- Four-step approach
  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 $p_0 \leq \mathbb{P}(y = 1 \mid x) \leq p_1$ (informally)
- Total of 8 model pipelines
- Hyperparameter tuning
  - Model parameters and K
  - Stratified 10-fold CV
  - Bayesian optimization with Optuna
  - Optimize for log loss

Bet Sizing

- Suppose an event has m fights
  - These are essentially fought back-to-back
  - We want to place all our bets at once
- 2^m different outcome sequences (ignoring draws, no contests)
- $2m + 1$ bets
  - One bet per fighter
  - A risk-free "no-bet" option that returns the full wager w.p. 1
- How should one allocate their money?
  - One option is to optimize for the growth rate of capital
  - This has some desirable properties (e.g., has an expected time to reach a specified goal that is asymptotically less than any other strategy)

Simultaneous ("Classical") Kelly

- For returns matrix $R \in \mathbb{R}^{(2m+1)\times 2^m}$ and outcome probabilities $\pi \in \Delta_{2^m}$, we want to find allocation $b \in \mathbb{R}^{2m+1}_{\geq 0}$ to maximize the expected log growth rate of wealth, $G_\pi(b) = \pi^T \log(R^T b)$:

$$\begin{aligned} \max_b \quad & \pi^T \log(R^T b) \\ \text{s.t.} \quad & 1^T b = 1 \\ & b \geq 0 \end{aligned}$$

- Ex. 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 $\Pi = \{\pi \in \Delta_{2^m} \mid A\pi \leq d\}$:

$$\begin{aligned} \max_{b,\lambda} \quad & \min(\log(R^T b) + A^T \lambda) - d^T \lambda \\ \text{s.t.} \quad & 1^T b = 1 \\ & b \geq 0 \\ & \lambda \geq 0 \end{aligned}$$

- Construct A and d as

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

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

- 8-year backtest period
  - Start of 2017 to end of 2024
  - 3960 bouts, 331 events
- Model pipelines are refit after every event, hyperparameters retuned at the end of each year
- Initial bankroll of $1000
- Closing odds from Bovada Sportsbook used to determine wagers
  - $0.50 minimum bet size $\implies$ all bets less than $0.50 rounded down to $0
  - Draws and no contests return full corresponding wager
- Model pipelines with Venn-Abers combined with both betting strategies
  - Total of 12 model pipeline and betting strategy combinations
  - Total of 36 combinations when considering different fractions

Monte Carlo Hypothesis Testing

- How do we know our observed profit is not by chance?
- Under null hypothesis, our approach has no edge $\iff$ closing odds reflect the true outcome probabilities
  1. Compute devigged implied probabilities (normalize to remove house edge)
  2. Sample outcome sequences according to these probabilities
  3. Compute profit/loss in this simulated reality
  4. Repeat a lot of times (10000)
- Calculate p-value as

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

  and adjust 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: Summary of model metrics over the backtest period, compared with closing odds

Model Results (cont.)

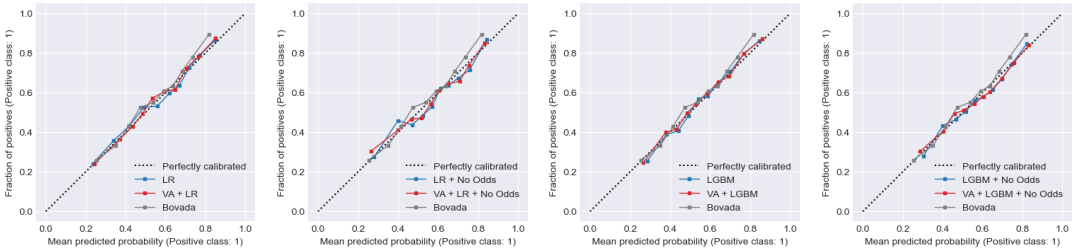


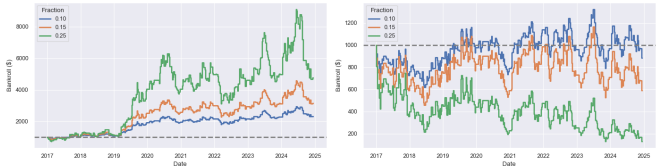
Figure: Calibration plots comparing model pipelines with and without Venn-Abers

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	
		0.15	2159.40	1196	5.25	-32.02	**0.0516**
		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
	Distrib. 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
	Distrib. 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
	Distrib. 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
	Distrib. Robust	0.10	-394.33	1380	-2.57	-62.51	1.0000

Table: Betting metrics by model pipeline and betting strategy combination

Betting Results (cont.)



(a) Logistic regression (b) Logistic regression, no odds

Figure: Bankroll over time for selected simultaneous Kelly strategy combinations

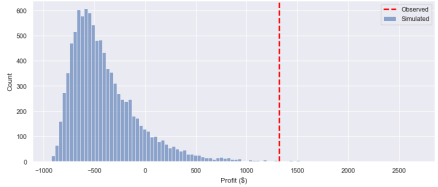


Figure: Simulated profit vs. observed, LR and simultaneous Kelly ($f = 0.10$)

Case Study: Logistic Regression and Simultaneous Kelly

- Clearly, logistic regression with simultaneous Kelly give the best results
  - Concluding with this is boring
- Further questions
  - What features are driving these results?
  - Why does most of our bankroll growth take place 2019 and 2020, but stagnate afterwards?
  - What types of fights is our model struggling with and why?
  - Can we identify a stricter subset of fights to further optimize profits?

Case Study (cont.)

- Opening odds-derived feature at top, but importantly does not dominate the model
- Majority of features in top 20 are derived from sources other than UFC Stats

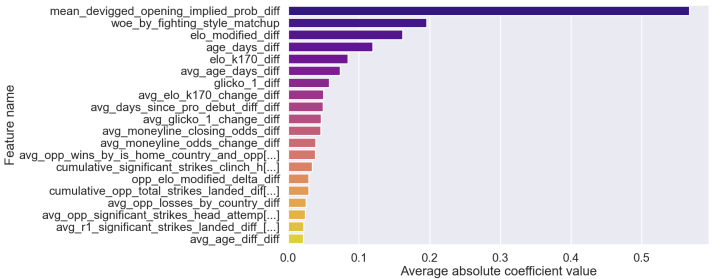


Figure: Top 20 features by average absolute coefficient value in logistic regression model

Case Study (cont.)

Year	Log Loss			Brier Score		
	Model	Bovada	Δ	Model	Bovada	Δ
2017	0.614738	0.609460	0.005277	0.212572	0.210317	0.002256
2018	0.597811	0.597080	0.000731	0.206408	0.205833	0.000575
2019	0.624346	0.636898	-0.012552	0.217742	0.223757	-0.006016
2020	0.611094	0.617461	-0.006367	0.211575	0.214045	-0.002470
2021	0.622634	0.621523	0.001110	0.217409	0.216309	0.001100
2022	0.599701	0.597913	0.001788	0.207473	0.205662	0.001811
2023	0.608926	0.602462	0.006465	0.210956	0.207674	0.003282
2024	0.586208	0.583386	0.002822	0.199778	0.198925	0.000853

Table: Model pipeline and sportsbook log loss and Brier score by year

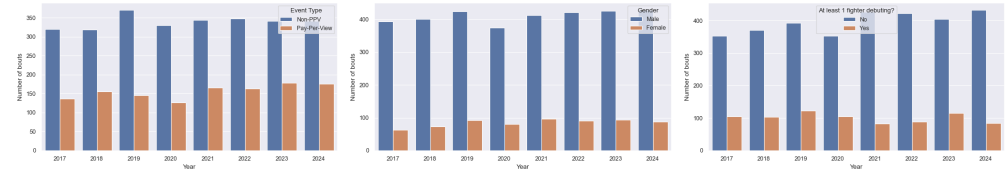


Figure: Number of bouts per year by event type, gender, and debuts

Case Study (cont.)

Event Type	Log Loss			Brier Score		
	Model	Bovada	Δ	Model	Bovada	Δ
Pay-Per-View	0.593012	0.596247	-0.003235	0.203459	0.204517	-0.001057
Non-PPV	0.614957	0.613594	0.001363	0.213644	0.212900	0.000744

Gender	Log Loss			Brier Score		
	Model	Bovada	Δ	Model	Bovada	Δ
Male	0.603392	0.606017	-0.002625	0.208329	0.209354	-0.001025
Female	0.630458	0.618340	0.012118	0.220585	0.214636	0.005949

UFC Experience	Log Loss			Brier Score		
	Model	Bovada	Δ	Model	Bovada	Δ
Both debuting	0.607946	0.579237	0.028709	0.210271	0.196018	0.014253
One fighter debuting	0.573801	0.560372	0.013429	0.195112	0.189045	0.006067
Both at least 1 fight	0.615511	0.619609	-0.004098	0.213782	0.215412	-0.001630
Both at least 3 fights	0.614461	0.617389	-0.002927	0.213161	0.214277	-0.001116
Both at least 5 fights	0.613856	0.617412	-0.003556	0.212762	0.214238	-0.001476

Table: Model pipeline and sportsbook log loss and Brier score by event type, gender, and experience

Case Study (cont.)

- What if we ignore women's divisions and debuts?
- Rerun everything except on subset of original data

	Log Loss	Brier Score
Logistic Regression	0.611250	0.211724
Bovada Sportsbook	0.616538	0.214093

Table: Model metrics over backtest period for fight subset, compared with closing odds

Fraction	Profit ($)	Total Bets	Yield (%)	MDD (%)	Raw p-value
0.10	4758.95	964	15.99	-20.56	
0.15	11306.91	974	14.21	-29.42	0.0001
0.25	43830.51	976	11.07	-46.47	

Table: Betting metrics by fraction, fight subset

Case Study (cont.)

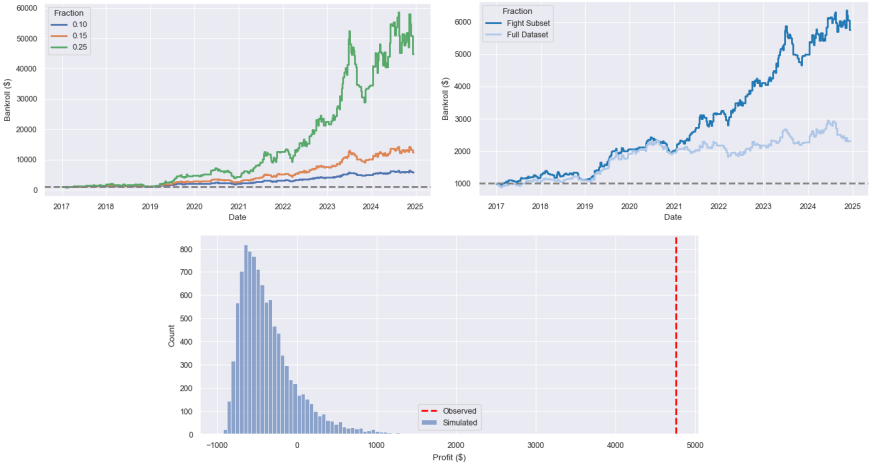


Figure: Selection of betting performance plots, fight subset

A Few Takeaways

- Sportsbooks are difficult to beat
- Clear value-add from integrating alternative signals
- Market information is valuable
- Possible edge decay?

Future Work

- Improvements to feature engineering and selection
  - Time-weighted averages, Bayesian smoothing, genetic algorithms, etc.
- Data source ablation studies
  - Massive practical limitations imposed by having so many interdependencies
  - Can we trim it down?
- Collect even more data
  - Theoretically possible to get timestamped odds data from FightOdds.io
- Use dataset to answer other research questions
  - What impact does elevation have on striking/grappling outputs?
  - How do markets react to news such as late replacements, weight misses, etc.?
  - To what extent does a fighter's ability to rehydrate and regain weight after weigh-ins affect their likelihood of winning?

Miscellaneous

- GitHub: `https://github.com/ehan03`
  - Repository: `https://github.com/ehan03/yale-senior-thesis`
- Tech stack
  - Languages: Mainly Python, feature engineering in raw SQL
  - Main libraries: Scrapy, Pandas, NumPy, Scikit-learn, LightGBM, Optuna, CVXPY, SQLite, SQLAlchemy, venn-abers (Ivan Petej), Seaborn