S&DS 625: Final Project Abstract

Eugene Han

2024-10-11

Abstract

Judging in mixed martial arts, especially in high-profile promotions like the Ultimate Fighting Championship (UFC), can be controversial. Although the Association of Boxing Commission's MMA committee has established a set of criteria for scoring fights, notably "effective striking/grappling," "effective aggressiveness," and "fighting area control," these measures are not standardized nor particularly well-defined and consequently left to interpretation. Using round-by-round performance statistics (e.g. strikes landed, takedowns, ground control time, etc.) and event/fighter metadata (e.g. event location and fighter nationalities) from historical UFC fights, I aim to model per-round judge scores to understand what variables influence scoring and how their impact varies across different judges due to personal preferences. This analysis could reveal which judges weigh factors like grappling more heavily than other judges or if scoring is influenced by fighters fighting in their home country, which could be due to a more reactive crowd or a desire to avoid criticism from home country fans.

Data

I plan to use data from the UFC Stats and MMA Decisions websites:

- **UFC Stats**: The website contains detailed round-by-round statistics for all past UFC fights, with metrics including significant strikes landed/attempted broken down by the targeted body part and position, submssion attempts, takedown success rate, and control time. To respect the UFC's Terms of Service, I will not be directly scraping their website and instead use the data aggregated in a public GitHub repository. The repository contains six CSV files containing event metadata, fight metadata, fight results, fight statistics by round, fighter names, and fighter metadata like height and date of birth. As a result, I already have the data for this source.

```r
# Read in fight results data
r <- read.csv("../data/UFC Stats/ufc_fight_results.csv")
head(r, 3)
```

```
##                          EVENT                              BOUT
## 1 UFC 307: Pereira vs. Rountree Jr.  Alex Pereira  vs. Khalil Rountree Jr.
## 2 UFC 307: Pereira vs. Rountree Jr.   Raquel Pennington  vs. Julianna Pena
## 3 UFC 307: Pereira vs. Rountree Jr.        Jose Aldo  vs. Mario Bautista
##   OUTCOME                      WEIGHTCLASS         METHOD ROUND TIME
## 1     W/L    UFC Light Heavyweight Title Bout         KO/TKO     4 4:32
## 2     L/W UFC Women's Bantamweight Title Bout Decision - Split     5 5:00
## 3     L/W                 Bantamweight Bout Decision - Split     3 5:00
```

```
##           TIME.FORMAT        REFEREE
## 1 5 Rnd (5-5-5-5-5) Marc Goddard
## 2 5 Rnd (5-5-5-5-5) Jason Herzog
## 3     3 Rnd (5-5-5) Mike Beltran
##                                                          DETAILS
## 1                                  Punch to Head At Distance
## 2 Sal D'amato 47 - 48.Derek Cleary 48 - 47.Mike Bell 47 - 48.
## 3   Chris Lee 29 - 28.Derek Cleary 28 - 29.Mike Bell 28 - 29.
##                                                              URL
## 1 http://ufcstats.com/fight-details/8c5e06c71f392f33
## 2 http://ufcstats.com/fight-details/1d68bb733de24b06
## 3 http://ufcstats.com/fight-details/3d7d82fff71e395a
```

```r
# Filter to fights that ended in decisions, follow the modern round structure,
# and has known scores
d <- r[which(grepl("^Decision", r$METHOD) &
             r$TIME.FORMAT %in% c("5 Rnd (5-5-5-5-5)", "3 Rnd (5-5-5)") &
             r$DETAILS != ""), ]

# Each round is scored by three judges
# Total of 3555 fights, 11185 rounds, and 33555 individual judge scores
# These represent upper bounds since we would need to merge with the data
# available from the second data source
cat(nrow(d), sum(d$ROUND), sum(d$ROUND) * 3)
```

```
## 3555 11185 33555
```

```r
# Example of round-by-round fight statistics for a recent fight that ended
# in a (controversial) decision
s <- read.csv("../data/UFC Stats/ufc_fight_stats.csv")
s[which(s$BOUT == "Jose Aldo vs. Mario Bautista"), -c(1, 2)]
```

```
##       ROUND        FIGHTER KD SIG.STR. SIG.STR... TOTAL.STR.      TD TD.. SUB.ATT
## 19 Round 1      Jose Aldo  0 13 of 26        50%    20 of 38 0 of 0  ---       0
## 20 Round 2      Jose Aldo  0 21 of 43        48%    25 of 50 0 of 0  ---       0
## 21 Round 3      Jose Aldo  0 17 of 48        35%    20 of 51 0 of 0  ---       0
## 22 Round 1 Mario Bautista  0 20 of 53        37%    40 of 81 0 of 2   0%       0
## 23 Round 2 Mario Bautista  0 19 of 53        35%    31 of 65 0 of 4   0%       0
## 24 Round 3 Mario Bautista  0 10 of 36        27%    19 of 50 0 of 4   0%       0
##    REV. CTRL    HEAD    BODY      LEG DISTANCE CLINCH GROUND
## 19    0 0:00  8 of 21 1 of 1  4 of 4 12 of 25 1 of 1 0 of 0
## 20    0 0:00 19 of 40 2 of 3  0 of 0 18 of 40 0 of 0 3 of 3
## 21    0 0:00 14 of 44 3 of 4  0 of 0 16 of 46 1 of 2 0 of 0
## 22    0 2:20  8 of 39 5 of 6  7 of 8 15 of 48 5 of 5 0 of 0
## 23    0 2:14 15 of 46 2 of 5  2 of 2 19 of 52 0 of 1 0 of 0
## 24    0 2:41  3 of 26 5 of 8  2 of 2 10 of 36 0 of 0 0 of 0
```

```r
# Clarifying some column names
# - "KD": Knockdowns scored
# - "SIG.STR.": Significant strikes landed vs. attempted
# - "SIG.STR...": "SIG.STR. %" in CSV; same as "SIG.STR." column but as percent
# - "TOTAL.STR.": Total strikes landed vs. attempted
# - "TD": Takedowns landed vs. attempted
```

```
#  -  "TD..": "TD" column as percent
#  -  "SUB.ATT": Submission attempts
#  -  "REV.": Reversals in grappling position
#  -  "CTRL": Control time on ground or against cage
#  -  "HEAD", "BODY", "LEG": Significant strikes landed vs. attempted by body part
#  -  "DISTANCE", "CLINCH", "GROUND": Significant strikes landed vs. attempted
#    by fighting position
```

- MMA Decisions: The website contains per-round judge scores, consensus fan-submitted scores, judge identities, and some fighter metadata for MMA fights (including but not limited to the UFC) that ended in decisions. Unlike the official UFC, this website does not have an explicit Terms of Service available and its robots.txt file disallows crawlers from accessing only the non-public/administrative parts of the website. I would need to scrape this website using a package like `rvest` and restrict to factual, non-personal information; the package `polite` could additionally be used to ensure responsible scraping. Here is a screenshot of the information available for the same recent fight, Jose Aldo vs. Mario Bautista: