S&DS 625 Final Project

Eugene Han

2024-12-11

Abstract

Judging in mixed martial arts (MMA), particularly within the UFC, is often contentious due to subjective interpretations of scoring criteria. This project examines the relationship between fight statistics and round-level scoring, emphasizing variability across judges. By merging data from MMA Decisions and UFC Stats, the study identifies significant predictors of score differences, including striking and grappling metrics, and highlights notable discrepancies in how judges evaluate these factors. Unlike prior predictive modeling efforts, this work focuses on inferential analysis at the individual judge level, offering insights into scoring biases and variability, thereby contributing to the understanding of subjectivity in MMA judging.

1 Introduction

Judging in mixed martial arts (MMA), especially in high-profile promotions like the Ultimate Fighting Championship (UFC), can be controversial. Although the Association of Boxing Commissions (ABC) MMA committee has established a set of criteria for scoring fights, notably "effective striking/grappling," "effective aggressiveness," and "fighting area control" in order of decreasing priority, the application of these guidelines is often subjective and open to judges' interpretations. By analyzing round-level scoring data coupled with corresponding fight performance statistics, this project aims to explore how different variables, such as the difference in strikes landed, influence scoring and how that impact may vary across judges.

The dataset used for this investigation is a combination of two separate datasets sourced from the websites MMA Decisions (scoring data) and UFC Stats (fight performance data). Prior work, such as Nate Latshaw's "Judge AI" project (Latshaw 2021), has taken a similar approach in terms of combining data from these same two websites but has primarily focused on predictive modeling applications to predict a "consensus" judge score by round. Our work differs in the sense that we are interested in inferential questions and scoring at the individual judge level, rather than a proxy "ground truth" score.

The report will be structured as follows: Section 2 will provide an overview of the data used in this project, including the process of obtaining, cleaning, and merging the two sources; Section 3 presents the results of data exploration and our analyses; and Section 4 will summarize our findings, limitations, and future scopes.

2 Data Overview

A significant part of this project was spent obtaining the final dataset due to challenges in scraping, cleaning, and merging the data. Before describing this process in detail, we will first briefly introduce the final dataset with descriptions and sample rows/columns. Below is one such sample for the most recent fight in our data, Rob Font vs. Kyler Phillips on October 19, 2024:

round	red_sig_strikes_landed	blue_sig_strikes_landed	judge_id	red_score	blue_score
1	7	10	357	9	10
1	7	10	318	9	10
1	7	10	127	9	10
2	31	14	357	10	9
2	31	14	318	10	9
2	31	14	127	10	9
3	23	14	357	10	9
3	23	14	318	10	9
3	23	14	127	10	9

In this example, the fight was 3 rounds long and we have scoring data from all 3 judges, giving a total of 9 rows. Fighters either come from the "red" or "blue" corners of the arena, so fighter-specific column names come with a corresponding "red" or "blue" prefix. For example, `red_sig_strikes_landed` represents the number of significant strikes the fighter in the red corner landed on their opponent (fighter in the blue corner) in that round, while `blue_score` represents the score given to the fighter in the blue corner by the judges. All judge scores use a 10-point scale described in the judging criteria/scoring document from the ABC. The `judge_id` column refers to a unique identifier associated with each judge that references the data scraped from MMA Decisions; here are the corresponding rows from the relevant dataframe:

id	name
127	Junichiro Kamijo
318	Derek Cleary
357	Michael Bell

It's important to note that we only displayed a small subset of the columns in the final dataset. The following is a more comprehensive list:

- `[ufcstats/mmadecisions]_bout_id` - Unique identifier for each fight
- `[ufcstats/mmadecisions]_event_id` - Unique identifier for the event that the fight belongs to
- `[red/blue]_[ufcstats/mmadecisions]_fighter_id` - Unique identifier for each fighter
- `round` - Round number, either 1 through 3 or 1 through 5
- `[red/blue]_knockdowns_scored` - Number of knockdowns by red/blue corner fighter on opponent due to strikes
- `[red/blue]_total_strikes_[landed/attempted]` - Number of total strikes landed/attempted by red/blue corner fighter
- `[red/blue]_sig_strikes_[landed/attempted]` - Number of significant strikes landed/attempted in total across all body regions and positions by red/blue corner fighter
- `[red/blue]_sig_strikes_[head/body/leg]_[landed/attempted]` - Number of significant strikes landed/attempted by red/blue corner fighter by opponent's body region
- `[red/blue]_sig_strikes_[distance/clinch/ground]_[landed_attempted]` - Number of significant strikes landed/attempted by red/blue corner fighter by position
- `[red/blue]_takedowns_[landed/attempted]` - Number of takedowns landed/attempted by red/blue corner fighter

- `[red/blue]_submissions_attempted` - Number of submission attempts by red/blue corner fighter
- `[red/blue]_reversals_scored` - Number of times red/blue corner fighter reversed their position in grappling or in the clinch from a disadvantageous position to a more advantageous one (e.g. bottom to top control)
- `[red/blue]_control_time_seconds` - Number of seconds red/blue corner fighter maintained and controlled an advantageous position in grappling

2.1 Acquisition

To respect the UFC's Terms of Service, we did not directly scrape fight data from the UFC Stats website and instead elected to use the data collected and aggregated in a public GitHub repository by user Greco1899 (https://github.com/Greco1899/scrape_ufc_stats). The data used for this report was downloaded on October 21, 2024 to ensure a fixed cutoff date since the repository updates on a weekly basis.

To obtain data from the MMA Decisions website, we utilized `rvest` for extracting relevant information from the HTML structure and `polite` for session management and crawling web pages respectfully, ensuring a 5 second delay between requests. The general strategy we took was to grab all necessary URLs and saving them to disk before extracting data and creating the final dataframes so that we avoid any excessive page revisits.

2.2 Cleaning

Cleaning the UFC Stats data mainly revolved around converting strings into numeric values and standardizing the data to rely on unique IDs over potentially non-unique names/strings. A particularly relevant example of converting strings to numbers were statistics involving successes and attempts. For instance, the scraped data might list "1 of 2" for takedowns which needs to be parsed into 1 takedown landed and 2 takedowns attempted. In terms of unique IDs, one of the biggest problems in the scraped data was how fight-level data was stored. While each fight could easily be traced back to a unique bout ID, the IDs of the two fighters were not stored by fight and instead a fight "name" was stored in the form "Red Fighter Name vs. Blue Fighter Name." This was especially problematic because some fighters have identical first and last names. Moreover, the GitHub user had wrote their scripts such that fighter data would only append new entries while all other data would be rewritten every week; any fighters whose names had been updated over time made it so that a simple join would not handle all cases. As a result, a nontrivial amount of the matching had to be done manually by cross-referencing the UFC Stats website. The motivation behind all of this was that if we could successfully create a one-to-one mapping between the event IDs and fighter IDs from UFC Stats and those of MMA Decisions, the fight IDs could then be matched trivially so that all fight statistics had corresponding judge scores.

Cleaning the MMA Decisions data was significantly easier as a lot of the preprocessing logic had been baked directly into the code for scraping the information from the website. Only the fighter data needed attention to handle duplicate fighters and parsing out fields like dates of birth and nicknames from free-form text which was an artifact of the poor HTML structure of fighter pages.

2.3 Merging

We first tried to match the events between the two data sources. We selected only the event IDs from UFC Stats for which at least one fight in the corresponding event ended in a decision based on the outcome method. We then made the assumption that both data sources had been scraped in such a way that respected their chronological order, making the matching a simple concatenation of event ID columns. To sanity check this, we computed the number of fights that ended in decision by event per data source and manually inspected any discrepancies. All discrepancies were confirmed to be edge cases where the fights had originally ended in decisions but were later overturned to no contests (due to reasons such as failed drug tests) according to UFC Stats, but were not listed on the MMA Decisions website.

Next, we matched the fighters. We first filtered for fighter IDs that were present only in fights that originally ended in decisions according to UFC Stats, narrowing the number of candidates and reducing the chance for false one-to-many relationships. We then filtered for fighters from both data sources that had unique names and temporarily saved the exact name matches. With the remaining fighters, we repeated this process for exact matches for unique dates of birth and nicknames and finally manually matching the remaining unmatched fighters.

To match bout IDs, we took the bout data for MMA Decisions, which had the bout ID, event ID, and both fighter IDs with respect to the MMA Decisions website, joined it with the event and fighter ID mappings mentioned above, and finally joined this result with the bout data for UFC Stats, which also contains the bout ID, event ID, and both fighter IDs with respect to the UFC Stats website. The last step was to join this with the fighting statistics from UFC Stats and judge scoring data from MMA Decisions, making sure the round numbers lined up.

2.4 Final Preprocessing

This merged dataset still had a few issues to address. First, judges can deduct one or more points in any round for certain infractions such as eye pokes or egregious fence holding. This can create situations where the fight statistics clearly point to a "10-9" round in favor of a fighter, but a point deduction results in a "9-9" round. Since these deductions are shared across all judges and completely independent of the striking and grappling statistics, we are more interested in investigating the "what-if" scores without these deductions. We filtered for fights with these strange scores where both fighters' scores were less than 10 or corresponded to outcome method text information from UFC Stats that suggested point deductions. We then cross-referenced MMA Decisions to find out the exact round, number of points deducted, and the corresponding fighter for each of these cases and manually created a CSV file that could be joined with our merged dataset from above to add those points back.

Next, we filtered out any remaining rows where judges scored a "10-7" round for a fighter (equivalently, where fighters' scores differed by more than 2). Not only are these incredibly rare (only 3 cases remained), but judges are generally discouraged from scoring rounds as "10-7" per the ABC's guidelines. Moreover, these guidelines' description of a "10-7" round heavily overlaps with that of a "10-8" round.

Additionally, MMA Decisions has recorded some judge scores where the judge's identity was unknown, leading to a missing judge ID. Lastly, we include rows where judges have scored at least a full fight, a minimum of 3 rounds, resulting in a final dataframe with 28690 rows.

3 Results

We now present the results of our study, which includes data exploration and analysis. In the exploration section, we derive candidates for response variables as well as engineer a basic set of features based on domain knowledge that we believe would fundamentally influence judge scoring. We then visualize these variables and their relationships between each other and the response. In the analysis section, we then fit two models: a baseline linear regression model on all of our observations to sanity check our engineered features, and a second linear regression on a subset (fight rounds that were scored by the top 10 judges with the most rounds scored) that includes interaction terms to try to tease out individual judge preferences.

3.1 Exploration

One way to define a response variable for analysis purposes is to calculate the difference between `red_score` and `blue_score`, resulting in `score_diff` which is a discrete, numeric variable. Inspecting the distribution of these score differences, we see that judges rarely give a tie score (equivalently a "10-10") with the bulk of scores resulting in a fighter having a one point advantage. We also see that the fighter in the red corner tends to win rounds more often in both the one and two point advantage scenarios.

```
##
##    -2    -1     0     1     2  <NA>
##   429 11890    21 15673   677     0
```

Alternatively, we can create a factor by concatenating the scores into a string in the form "{red_score}-{blue_score}" and specify the order of the levels, resulting in an ordinal variable. This not only respects the natural ordering of score_diff but also accounts for the fact that the "gaps" between a "10-10" and a "10-9" versus between a "10-9" and "10-8" are not the same nor well-defined, which isn't captured by a discrete numeric scale. We can see that this preserves the distribution seen above.

```
##
##  8-10  9-10 10-10  10-9  10-8  <NA>
##   429 11890    21 15673   677     0
```

In total, there are 394 unique judges represented in this dataset, with significant variability in the number of scored rounds by judge, ranging from just 3 rounds to 2935 rounds.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00   12.00   72.82   32.00 2935.00
```

If we join with the judges data to use judge names over their IDs, we can obtain the top 10 judges with the largest number of rounds scored:

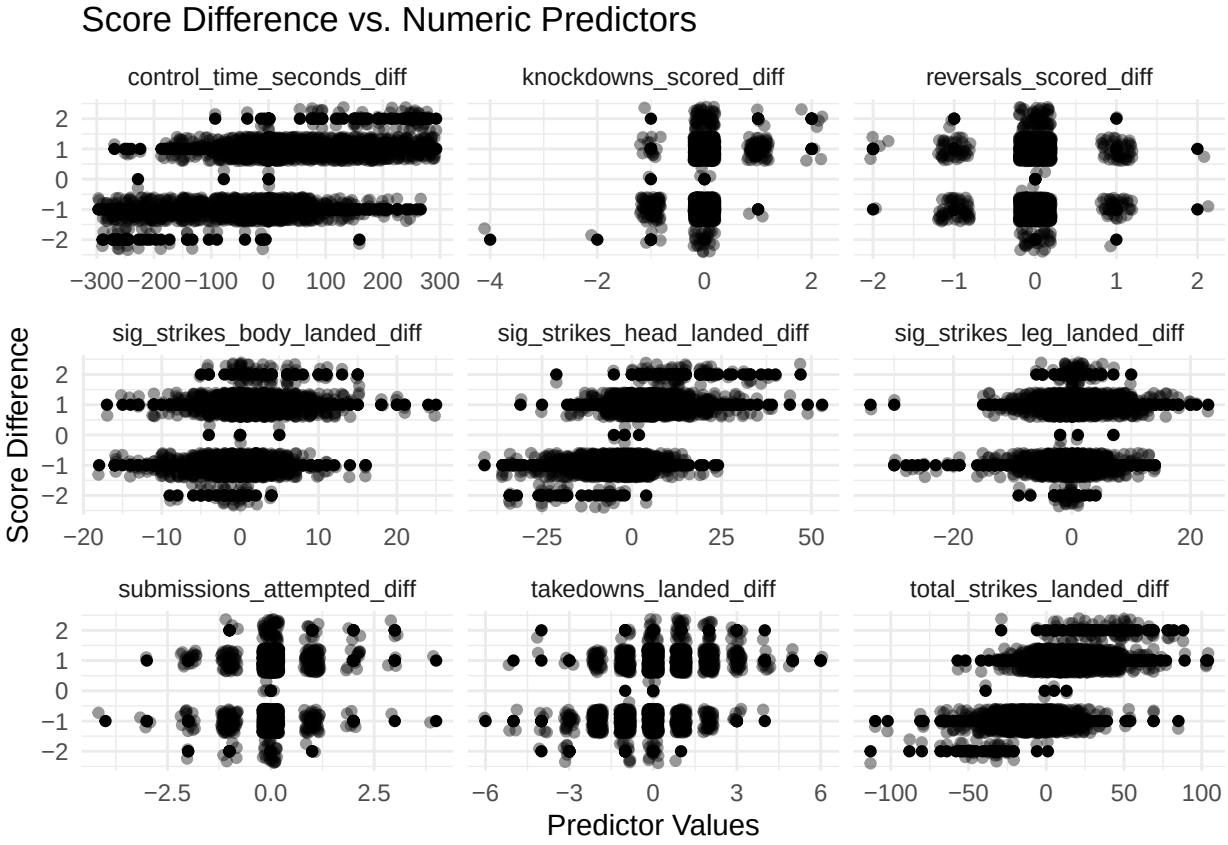
```
##          judge_name    n
## 1       Sal D'Amato 2935
## 2      Derek Cleary 1953
## 3         Chris Lee 1786
## 4      Michael Bell 1367
## 5  Junichiro Kamijo 1293
## 6        Eric Colón  951
## 7        Tony Weeks  947
## 8      Ron McCarthy  646
## 9     Adalaide Byrd  642
## 10    Ben Cartlidge  617
```

Next, we define 9 engineered features that encode comparative information that directly pit the two fighters' performance against each other.

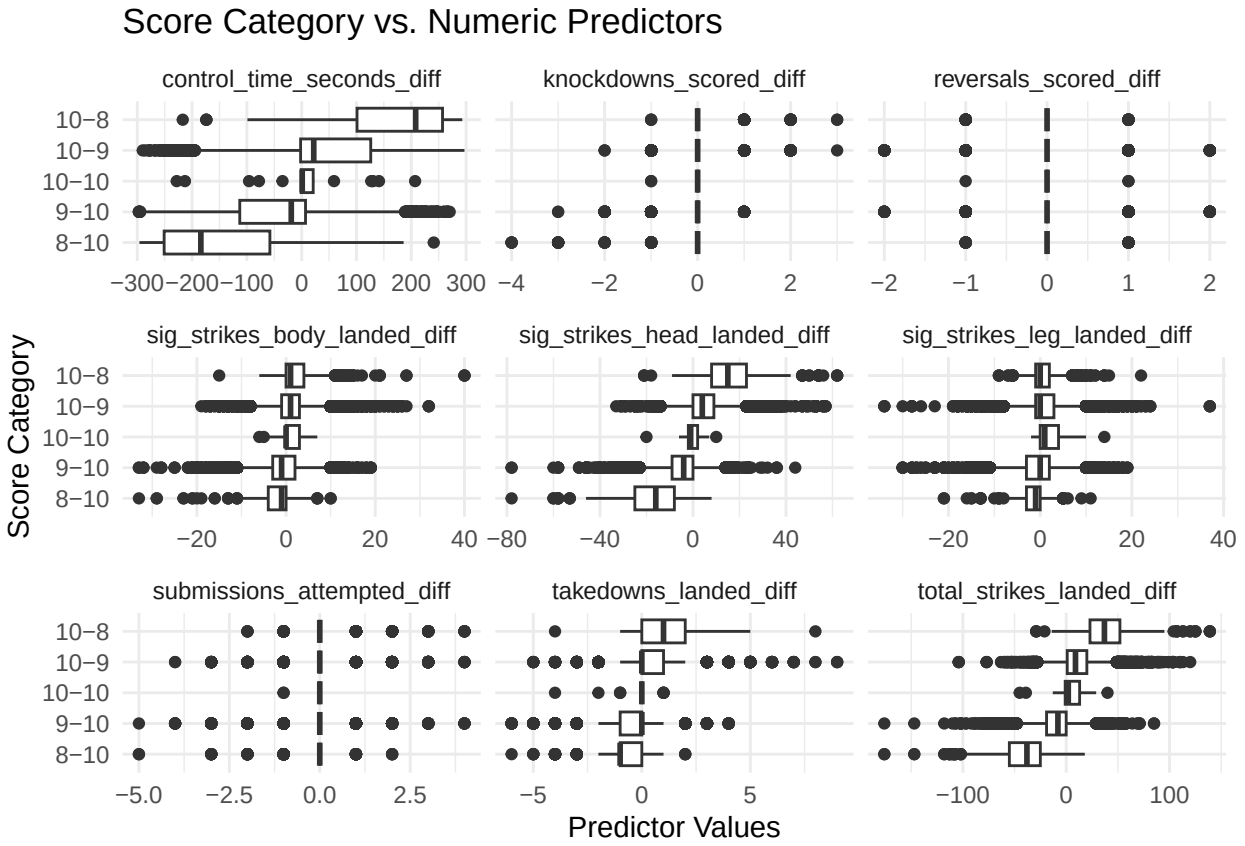
- knockdowns_scored_diff - Defined as (red_knockdowns_scored - blue_knockdowns_scored). If a fighter scores a knockdown, this typically means that they dealt a substantial amount of damage to their opponent in a strike, so we should expect a positive difference to be associated with a larger score discrepancy in favor of the fighter in the red corner.
- sig_strikes_[head/body/leg]_landed_diff - Defined as (red_sig_strikes_[head/body/leg]_landed - blue_sig_strikes_[head/body/leg]_landed). We break this feature down by the targeted body region since we should expect significant strikes to the head and body will be more impactful and damaging than those to the leg. We don't include the features by position since there would be some redundancy and there isn't a strong justification for why strikes from a certain position would contribute more to scoring than other positions.
- total_strikes_landed_diff - Defined as (red_total_strikes_landed - blue_total_strikes_landed). Captures any striking advantaage holistically, including non-significant strikes.

- `takedowns_landed_diff` - Defined as (`red_takedowns_landed` - `blue_takedowns_landed`). A large difference in the number of takedowns landed would suggest more aggressiveness and a dominance on the ground from a fighter.
- `submissions_attempted_diff` - Defined as (`red_submissions_attempted` - `blue_submissions_attempted`). A large difference in this variable would represent a higher skill in grappling and putting the opponent into disadvantageous situations.
- `reversals_scored_diff` - Defined as (`red_reversals_scored` - `blue_reversals_scored`). Similar justification as above.
- `control_time_seconds_diff` - Defined as (`red_control_time_seconds` - `blue_control_time_seconds`). Similar justification as above.

To visualize these features' relationship with the score difference, we plot scatterplots for each feature with `score_diff` for now. We sample 2500 rows for computational efficiency and to make the plot a bit cleaner since we are just interested in seeing if there are any general trends. Since there's quite a bit of overlapping points, we also jitter the points a little bit.



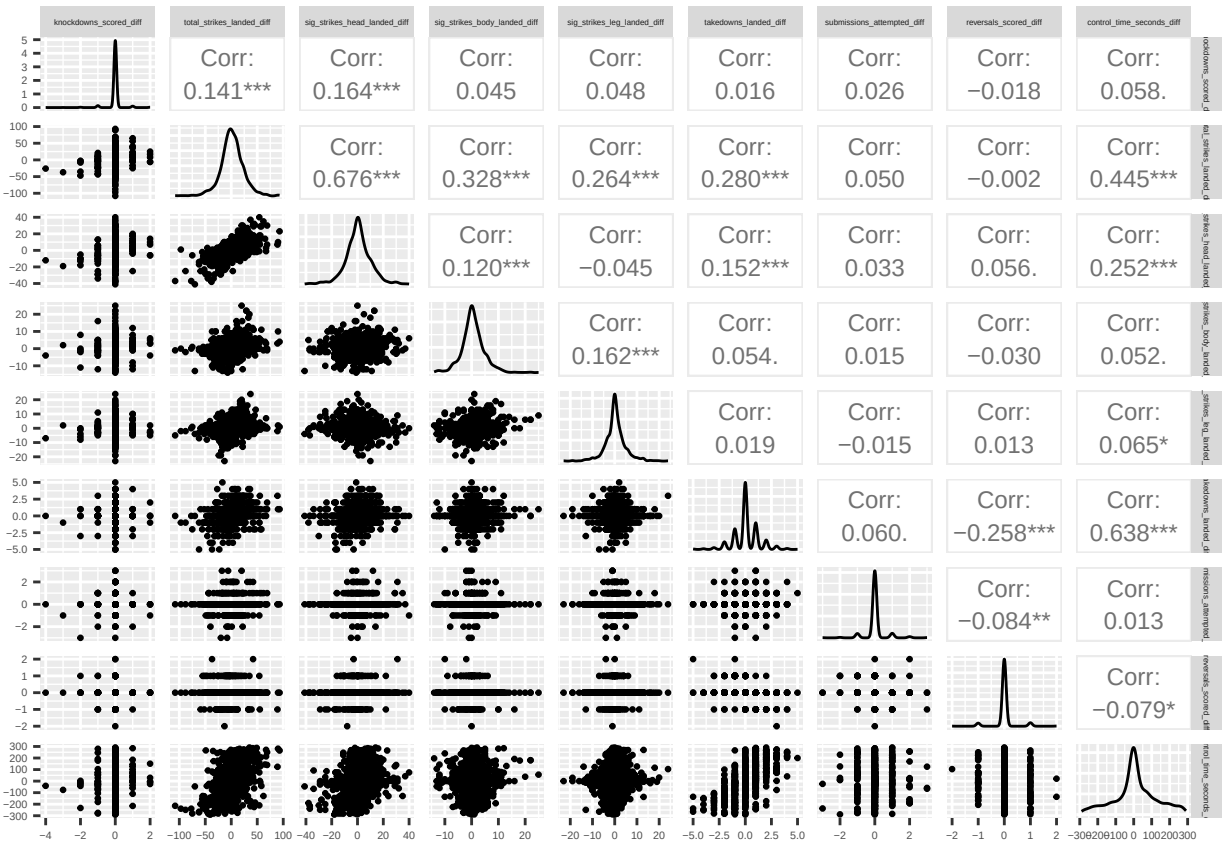
In general, we see approximately linear positive relationships between our engineered features and `score_diff` as hypothesized, although this isn't as clear for `reversals_scored_diff` and `submissions_attempted_diff`. We can also repeat these with boxplots and `score_string` using the full dataset.



One interesting observation we get here is that for `control_time_seconds_diff`, in order to score a "10-8"/"8-10" a fighter needs to have a substantial median control time advantage when compared to scoring a "10-9"/"9-10". The change between the median control time difference when going from a "10-10" score to "10-9"/"9-10" is much smaller than when going from "10-9"/"9-10" to "10-8"/"8-10".

Lastly, we make a pair plot of all the numeric predictors we constructed to investigate if there is any substantial multicollinearity. We only sampled 1000 rows to reduce the computational load of plotting. For the majority of the pairwise comparisons, our variables have little to no correlation ranging between around -0.3 to 0.3. With that being said, there are two pairs that stand out: `total_strikes_landed_diff` with `sig_strikes_head_landed_diff` and `takedowns_landed_diff` with `control_time_seconds_diff`. These pairs of variables have somewhat moderately strong correlations of 0.676 and 0.638, respectively.

This isn't very surprising since we expect fighters to aim more for the head to inflict more damage, so strikes to the head would naturally make up a larger portion of total strikes landed and scale accordingly. Moreover, if a fighter lands more takedowns, we should expect them to spend more time on the ground in a dominant position. Although this multicollinearity may cause issues with respect to having less reliable coefficient estimates, we believe it's still important to include these features since they capture different facets of performance that aren't necessarily redundant. For instance, if two fighters are neck and neck on significant strikes, any differences in overall total strikes may be a deciding factor. Also, takedowns landed measure a fighter's ability to get their opponent to the ground, but the control time quantifies how effective that fighter is on the ground and how well they take advantage of that positioning.



3.2 Analysis

3.2.1 Baseline Linear Regression

As a preliminary analysis, we fit a linear regression model with `score_diff` as our response variable, loosely treating it as a continuous variable, with our 9 engineered features as numeric predictors and `judge_name` as a categorical predictor. Since `judge_name` is a factor that hasn't been releveled, the default (alphabetical) ordering is used such that the judge "Aaron Chatfield" is our reference level.

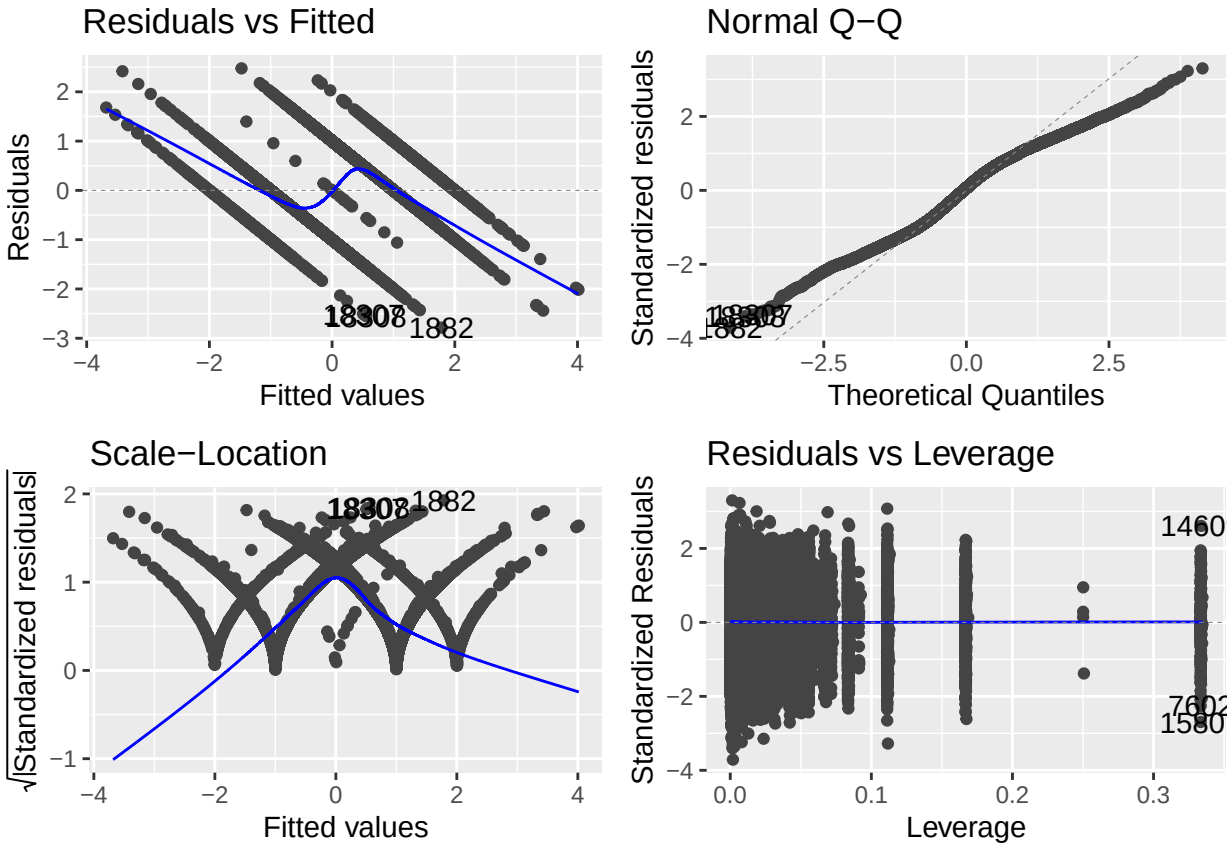
```
##
## Call:
## lm(formula = score_diff ~ knockdowns_scored_diff + total_strikes_landed_diff +
##     sig_strikes_head_landed_diff + sig_strikes_body_landed_diff +
##     sig_strikes_leg_landed_diff + takedowns_landed_diff + submissions_attempted_diff +
##     reversals_scored_diff + control_time_seconds_diff + judge_name,
##     data = xf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7847 -0.6132  0.0490  0.6053  2.4760
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.303e-01  1.120e-01   1.164  0.24460
## knockdowns_scored_diff          3.940e-01  1.589e-02  24.793  < 2e-16 ***
## total_strikes_landed_diff       4.733e-03  3.509e-04  13.488  < 2e-16 ***
```

```
## sig_strikes_head_landed_diff      3.582e-02  6.416e-04  55.837  < 2e-16 ***
## sig_strikes_body_landed_diff      2.693e-02  1.086e-03  24.801  < 2e-16 ***
## sig_strikes_leg_landed_diff       2.433e-02  1.056e-03  23.039  < 2e-16 ***
## takedowns_landed_diff             9.306e-02  4.903e-03  18.980  < 2e-16 ***
## submissions_attempted_diff        1.654e-01  8.626e-03  19.169  < 2e-16 ***
## reversals_scored_diff             7.276e-02  1.545e-02   4.710 2.49e-06 ***
## control_time_seconds_diff         2.199e-03  5.457e-05  40.291  < 2e-16 ***
## judge_nameAaron Menard           -1.345e-02  4.478e-01  -0.030  0.97605
## judge_nameAbe Belardo             1.460e-01  2.095e-01   0.697  0.48584
## judge_nameAdalaide Byrd          -3.135e-02  1.158e-01  -0.271  0.78663
## judge_nameAdrian Castro          -6.530e-01  2.441e-01  -2.676  0.00746 **
##  [ reached getOption("max.print") -- omitted 389 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7509 on 28287 degrees of freedom
## Multiple R-squared:  0.4912, Adjusted R-squared:  0.484
## F-statistic: 67.93 on 402 and 28287 DF,  p-value: < 2.2e-16
```

Based on our model summary, our baseline approach has multiple and adjusted R-squared values of 0.4912 and 0.484, respectively, indicating that it captures a moderate amount of the variance in `score_diff` and is overall a decent fit to our data. We see that all of our engineered numeric predictors are statistically significant, although this is partially an artifact of having a large sample size such that even small deviations from zero can be flagged as significant. Still, this model yields some useful information to us. We can see that the signs and magnitudes of the coefficients for each of these features align with our expectations that a larger positive value should be associated with scoring that increasingly favors the fighter in the red corner and that features corresponding to rarer but more impactful events such as knockdowns have larger coefficients.

For instance, when holding all other variables constant, an increase in `total_strikes_landed_diff` by 1 (which corresponds to the fighter in the red corner landing one more additional strike on their opponent) is associated with an increase in the score difference between red and blue fighters by about 0.005 in favor of red. Since strike counts can rack up quickly during a round, this makes sense; only a total dominance via a substantial gap in strikes landed would warrant consideration for scoring a round in favor of a "10-8" over a "10-9." On the other hand, an increase in `knockdowns_scored_diff` by 1 is associated with an increase in `score_diff` by 0.394 when all other predictors are held fixed. Knockdowns are infrequent and only occur when a fighter lands a particularly devastating strike, and so it is not surprising that this coefficient is so large relative to that of other predictors.

With that being said, it's important to point out the obvious limitations of this model. In particular, our response `score_diff` is discrete and bounded such that it only takes on 5 unique values. As a result, our model will predict unrealistic values given data that are either not integers and/or fall outside of the interval [-2, 2]. Moreover, we also observe strange behavior in some of our diagnostic plots due to the nature of `score_diff`.



Because the response variable is always one of {-2, -1, 0, 1, 2}, we see a series of "stripes" in our residuals vs. fitted plot as well as a collection of "V" shaped curves in our scale-location plot for each of the 5 values, which may indicate heteroskedasticity of residuals. There is some evidence of non-normality based on the Q-Q plot especially at the tails, but this isn't a huge surprise given the quantity of data we have. Based on the residuals vs. leverage plot, we don't see any observations that fall outside of Cook's distance values of 0.5 or 1 which suggests that we don't have any influential points.

3.2.2 Linear Regression with Interaction Terms

Next, in an attempt to estimate judge-specific coefficients of our predictors, we build on the previous model by including interaction terms between `judge_name` and each of the 9 engineered features. Because there are 394 unique judges in the full dataset, fitting such a model is impractical and computationally intensive. Consequently, we decided to take a subset of our dataset and filter for rounds that were scored by the top 20 judges with respect to the number of observations present. This leaves us with 17732 rows of data or about 62% of the original data with all judges having at least 320 observations, which will help us avoid egregiously large standard errors.

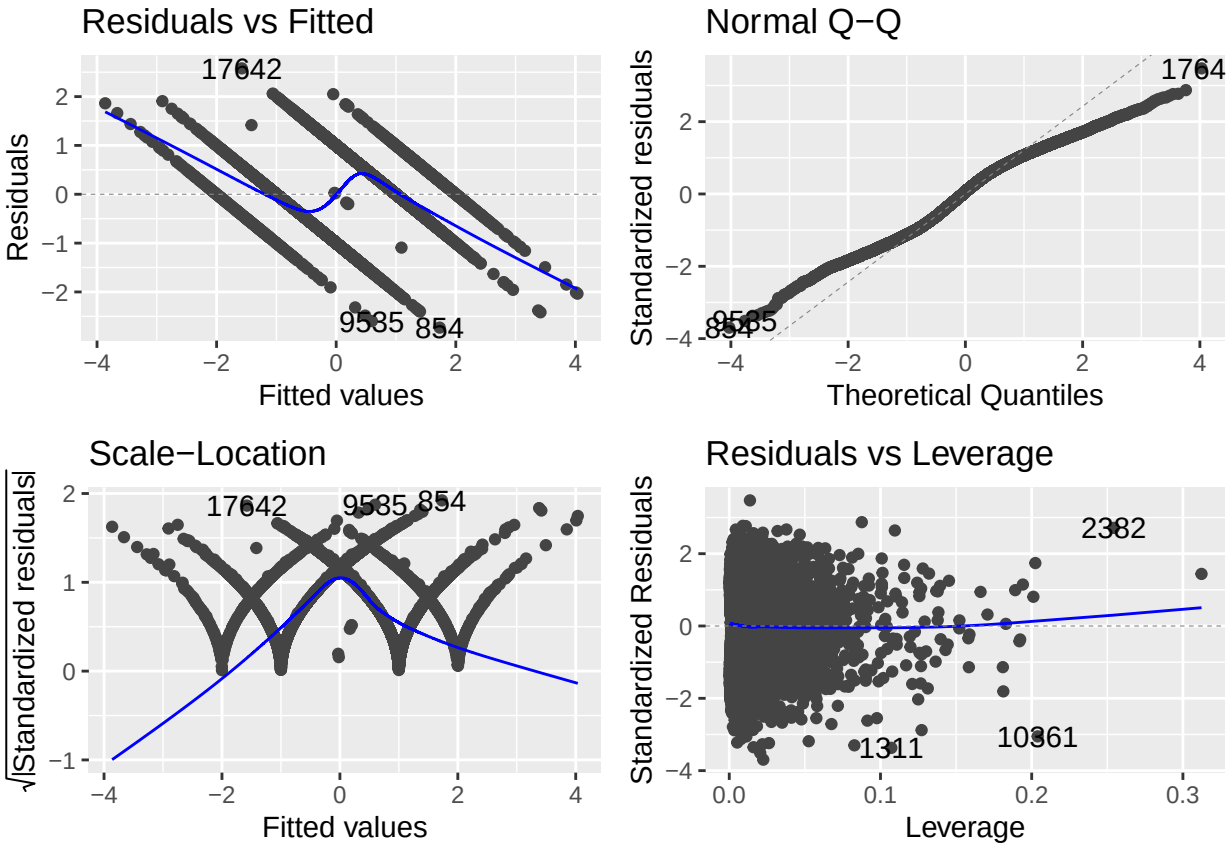
Since we want to view the predictors from the lens of each judge, we are most interested in estimating *marginal* effects. To avoid having to tediously compute these marginal effects by adding the interaction effects with the coefficients of the reference level as well as having to work with the covariance matrix to calculate standard errors, we use the shorthand `y ~ f / x` when specifying our model which will directly give us estimates of the marginal effects and the corresponding standard errors (McDermott 2019).

```
##
## Call:
## lm(formula = score_diff ~ judge_name/knockdowns_scored_diff +
##     judge_name/total_strikes_landed_diff + judge_name/sig_strikes_head_landed_diff +
```

```
##     judge_name/sig_strikes_body_landed_diff + judge_name/sig_strikes_leg_landed_diff +
##     judge_name/takedowns_landed_diff + judge_name/submissions_attempted_diff +
##     judge_name/reversals_scored_diff + judge_name/control_time_seconds_diff,
##     data = xf_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7306 -0.6119  0.0481  0.6062  2.5780
##
## Coefficients:
##                                                      Estimate Std. Error
## (Intercept)                                         0.1103996  0.0304821
## judge_nameBen Cartlidge                            -0.0520859  0.0435387
## judge_nameCardo Urso                               -0.1033151  0.0523594
## judge_nameChris Lee                                -0.0360081  0.0353026
## judge_nameDave Hagen                               -0.0843192  0.0456150
## judge_nameDave Tirelli                             -0.0522394  0.0482442
##                                                    t value Pr(>|t|)
## (Intercept)                                          3.622 0.000293 ***
## judge_nameBen Cartlidge                             -1.196 0.231590
## judge_nameCardo Urso                                -1.973 0.048490 *
## judge_nameChris Lee                                 -1.020 0.307750
## judge_nameDave Hagen                                -1.848 0.064547 .
## judge_nameDave Tirelli                              -1.083 0.278906
##  [ reached getOption("max.print") -- omitted 194 rows ]
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7472 on 17532 degrees of freedom
## Multiple R-squared:  0.5018, Adjusted R-squared:  0.4961
## F-statistic: 88.72 on 199 and 17532 DF,  p-value: < 2.2e-16
```

For brevity, we've truncated the summary output. This augmented model achieves multiple and adjusted R-squared values of 0.5018 and 0.4961, respectively, which are similar to our baseline and again indicates an decent fit to the data. Although not visible above, the estimates for the marginal effects correspond to the coefficients of the form `judge_nameAdalaide Byrd:knockdowns_scored_diff`. We will extract these out separately and visualize them later.

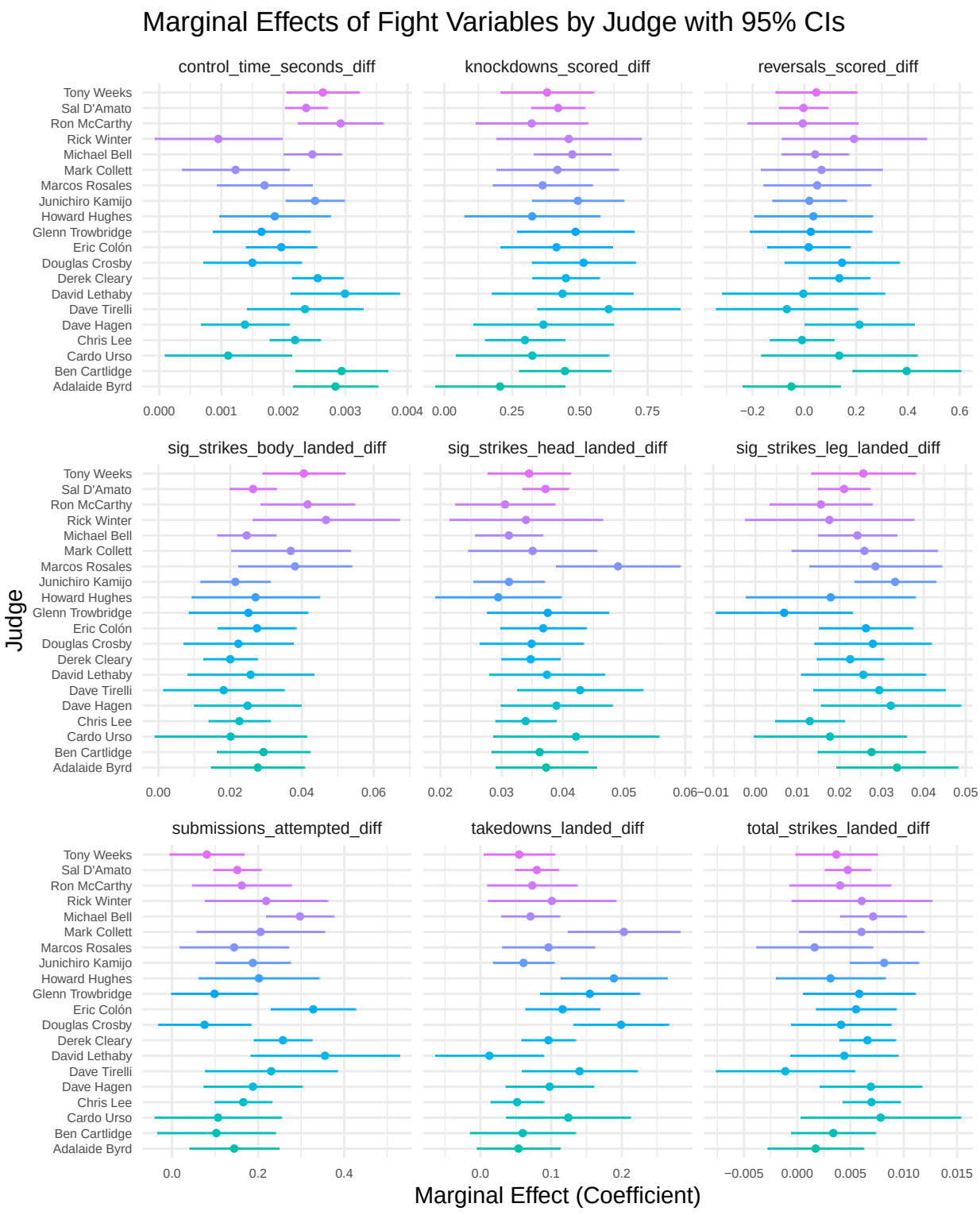
For completeness, we again generate the model diagnostic plots and see very similar outputs to those of the baseline model since we still have the issue of having a discrete response that takes one of five unique values.



Next, we can use the `tidy` function from `broom` to extract out and tabularize the model coefficients along with the corresponding lower and upper endpoints of the 95% confidence intervals. We then filter for the marginal effects by looking for the presence of a `:` symbol and extract out the judge name and feature name. The first several rows of this data frame are shown below.

Judge	Variable	estimate	conf.low	conf.high
Adalaide Byrd	knockdowns_scored_diff	0.2055145	-0.0331784	0.4442074
Ben Cartlidge	knockdowns_scored_diff	0.4448516	0.2760440	0.6136592
Cardo Urso	knockdowns_scored_diff	0.3244472	0.0424534	0.6064410
Chris Lee	knockdowns_scored_diff	0.2973803	0.1503288	0.4444317
Dave Hagen	knockdowns_scored_diff	0.3654661	0.1070833	0.6238488
Dave Tirelli	knockdowns_scored_diff	0.6064227	0.3432479	0.8695975

Using this, we can visualize the marginal effects by judge and facet by each predictor.



Based on the plots above, there appears to be some evidence of variability across judges within each of the predictors, and the extent of this variability seems to differ depending on the predictor. For example, the predictor `sig_strikes_head_landed_diff` is statistically significant for all judges at the 5% level based on the confidence intervals and are positive which aligns with our expectations. Moreover, the coefficient estimates are similar across the judges, with the exception of Marcos Rosales who seems to place more weight on significant head strikes; however, the confidence interval for Marcos Rosales overlaps with that of

many other judges, suggesting that this difference may not really be significant. On the other hand, there is substantially more variability in the marginal effects for `takedowns_landed_diff` across judges, with some like Adalaide Byrd, Ben Cartlidge, and David Lethaby showing evidence of differences in takedowns landed not being significantly different from zero. There also may be pairwise statistical differences between judges such as Douglas Crosby and Chris Lee, though this would need more rigorous testing and corrections for multiple comparisons. A similar takeaway can be made for `control_time_seconds_diff`, where there almost seems to be a clustering of judges who value control time dominance more than others. This is rather interesting as there has long been a debate among MMA fans whether or not a fighter who is simply "controlling" a top position in grappling without any accompanying effective strikes should be rewarded on the judge's scorecard.

Broadly, it seems like the marginal effects corresponding to striking-related variables (`knockdowns_scored_diff`, `sig_strikes_head_landed_diff`, `sig_strikes_body_landed_diff`, `sig_strikes_leg_landed_diff`, and `total_strikes_landed_diff`) are more consistent across judges, whereas those of grappling-related variables (`control_time_seconds_diff`, `reversals_scored_diff`, `submissions_attempted_diff`, and `takedowns_landed_diff`) show more variability. A potential explanation for this may be due to the black-and-white nature of striking, since it is usually visually obvious how much impact a strike has on a fighter, either by looking at damage on the face or a clear physical indication of pain or discomfort. On the other hand, grappling is a world of subtleties and may be more open to personal interpretation with respect to how much the tide of a fight shifts with respect to those variables.

Since our data only contains judging information for fights that went the distance, we have no information on how judges' partial scorecards for fights that may have gone a few rounds before ending due to a knockout, submission, or medical injury. As a result, the variability that we are seeing in the marginal effects may be influenced by this sort of natural "selection" bias. This could especially be true with respect to what we observed in the grappling-specific variables, as some fights can be heavily grappling due to a matchup between one or more fighters with wrestling backgrounds, while others can be pure striking between kickboxers; in both of these cases, there will always be some kind of striking, but that isn't necessarily true for grappling.

4 Conclusion

In this case study, we examined how judges score rounds in UFC matches by analyzing scoring data and fight statistics from multiple sources. The project involved extensive data acquisition, cleaning, and merging to create a comprehensive dataset, followed by exploratory and inferential analyses. Key findings include the identification of significant relationships between striking metrics and scoring patterns, as well as notable variability in how judges weigh grappling metrics such as takedowns and control time. This variability highlights the subjective nature of MMA judging, particularly for scenarios with either extensive or limited grappling.

There are some interesting and natural extensions of this work that could be the basis of future research. For example, one could perform an even deeper dive into the analysis of variability across judges, testing for pairwise differences and connecting those findings regarding judge preferences to past fights by actually re-watching some of the fights, especially ones with controversial decisions or specific rounds where judges heavily disagreed. One could also augment our dataset by also considering partially scored fights such as those found on the official UFC website ("UFC Scorecards" 2024), although it isn't clear how one would systematically collect this information since the scorecards are presented as images.

A major limitation of our analyses was the fact that the response variable `score_diff` was discrete and only took on 5 different values. Moreover, it isn't clear if scoring can be represented on a discrete scale with uniform gaps, since the conditions for scoring "10-8" rounds often requires a more stringent definition of "dominance." As such, it may be more appropriate to model the response as an ordinal variable and use something like an ordinal logistic or probit regression model, though at the cost of interpretability. Moreover, our data contains repeated measures since the values of our predictors are the same across all judges within the same round of the same fight, violating i.i.d. assumptions. As such, a hierarchical approach may also be a direction for investigation, whether that be through mixed effects or Bayesian modeling.

References

Association of Boxing Commissions. 2017. "MMA Judging Criteria/Scoring - Approved August 2, 2016." https://www.abcboxing.com/wp-content/uploads/2017/10/2017-Official-MMA-Judging-Criteria.pdf.

Greco1899. 2024. "Scrape_ufc_stats." https://github.com/Greco1899/scrape_ufc_stats.

Latshaw, Nate. 2021. "Introducing Judge AI." *LiteralFightNerd*. https://literalfightnerd.com/posts/2021-02-19-introducing-judgeai/.

McDermott, Grant. 2019. "Marginal Effects and Interaction Terms." https://grantmcdermott.com/interaction-effects/.

"MMA Decisions." 2024. https://mmadecisions.com/.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

"UFC Scorecards." 2024. https://www.ufc.com/scorecards.

"UFC Stats." 2024. http://www.ufcstats.com/statistics/events/completed.