Predicting UFC Fight Results

2023-12-11

Introduction

The UFC began as a professional mixed martial arts organization in 1993 serving as an alternative hand-to-hand combat sport that combined traditional boxing with wrestling, karate, kickboxing, and jiu-jistu fighting techniques. The entity was acquired by a group led by Dana White in 2001, who has served as the President for over two decades. Since estabilshing control, Dana White has exponentially grown the reach of UFC's product while also creating more structure and sanction to the sport of MMA. UFC currently has over 60 global broadcasting partners and is able to be accessed in over 165 different countries. With a traveling, tour-like model, the UFC has been able to sell-out many arenas across the world as equally become a highly-touted event to attend similar to boxing matches with well-known participants involved.

In many other North-american based sports, organizations have invested and founded their own analytics departments. These departments are responsible for using data to acquire and develop the right talent that will lead to on-field success and improve the team's product. Since the UFC's participants are individual fighters that often follow their own training regiment, there is a smaller focus on analytics within the sport.

The purpose of our project is two-fold. We want to evaluate fighters' historical data to determine fight styles that may posses a stronger correlation to sucess within the octagon. Identifying important factors will allow UFC fighters and their hired trainers to optimize their training regime, and will also benefit commentators in pointing out facets of the match the audience should keep in mind while spectating. Additionally, we want to create a model that maximizes predictive accuracy for the purposes of assisting sports bettors in finding potential opportunities for value not seen by the public. Dana White and the UFC have fully embraced the recent popularity of sports betting, forming sponsorships with companies such as bet365 in the UK, DraftKings in the US, as well as many others located worldwide. There is an established market for sports betting in the UFC, and we hope to create a model that provides an estimation of a winner between two fighters along with some form of uncertainty that allows a sports-better to determine if the predicted odds over or under-estimate a fighter's chance of winning compared to the sportsbook odds given to the public.

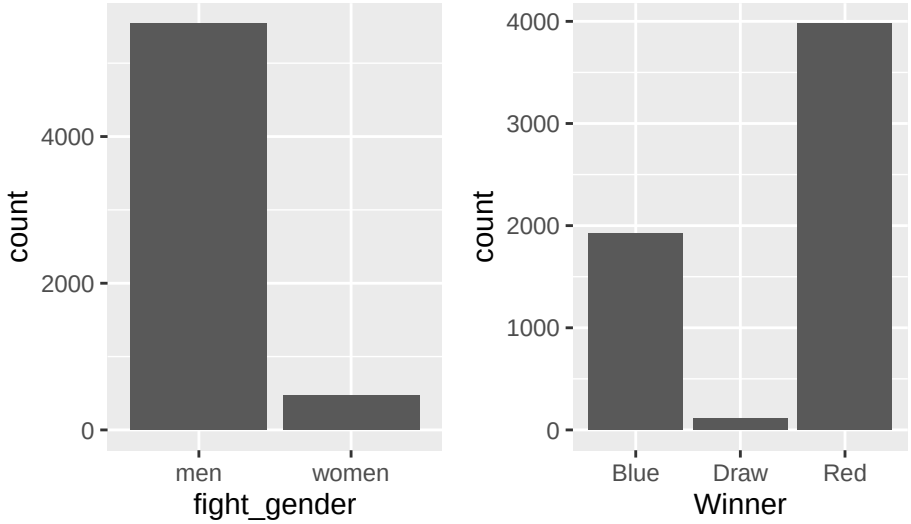
Our data is each UFC fight from 1994-2021, containing each fighter's names, physical information (age, height, weight), the amount of wins and losses in their UFC history, and various fighting data. The fighting data includes the average amount of attempted and landed attacks over their UFC career, as well as the frequency of different types of attacks they have faced from their previous opponents. We plan to fit an elastic net model that incorporates the standardization of a ridge regression model, and the variable selection of a lasso regression model, to determine which predictors in our dataset are most influential. Our response variable will be the winner in each fight, with that value randomized dependent on the color of the corner assigned during the fight (red or blue). The elastic net model will be best served for UFC commentators and trainers interested in how prior fighting strategies can lead to success in the future. The estimated probability values can be compared against the moneyline odds to determine estimated value for sports bettors, and the significant coefficients can be used by trainers and coaches to adjust their fighter's strategies.

Data Description and EDA

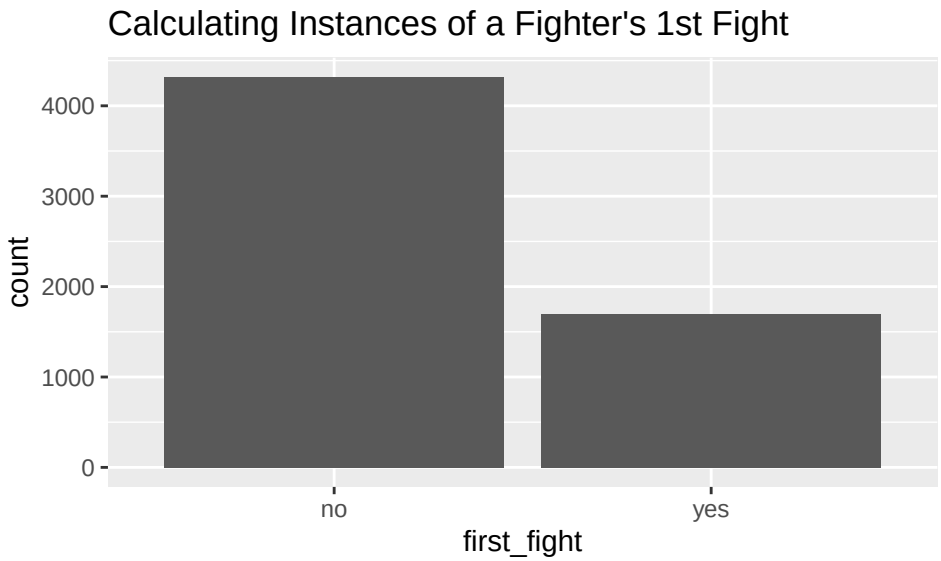
The dataset used for modeling was initially sourced from ufcstats.com, where the data was processed and published on kaggle.com. The dataset contains roughly 6,012 unique fights over a 27 year span, with each row containing data on fighters in the red and blue corner, a universal classification system used throughout

all rows. Since a given fighter could be in the red corner for one fight and in the blue corner for another, we don't place any value in the classification. The historical data for a given strike is split into four different columns. There is a separate column for the average amount of attempt and the average amount of strikes landed for a fighter's prior fight. Additionally, for that given strike there includes identical metrics for the averages of their prior opponents. For example, if Conor McGregor had one prior fight in which his opponent attempted 20 strikes to Conor's head and connected on 8, the row for Conor's second fight would display avg_opp_HEAD_att equal to 20 and avg_opp_HEAD_landed to 8.

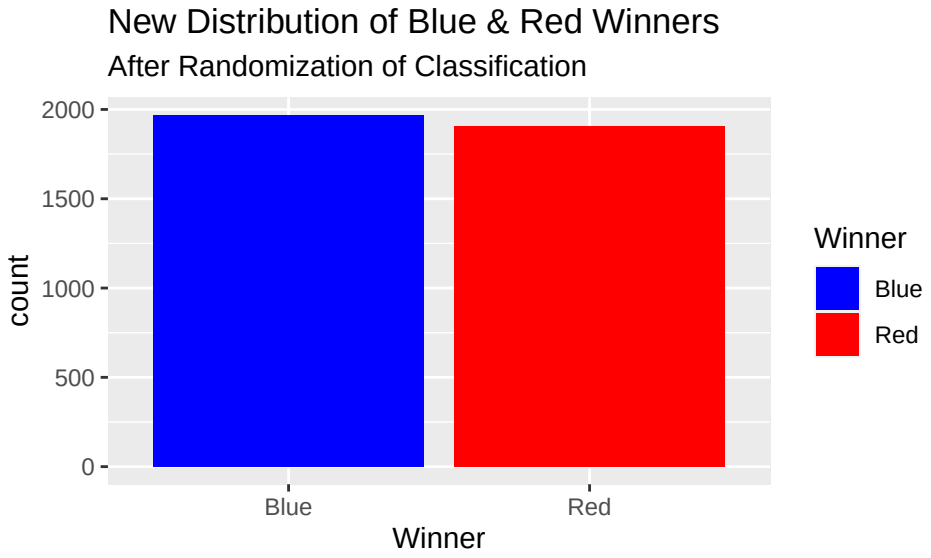
In the UFC, fighters can win in a variety of different ways. If both fighters are still standing at the end of the fighter, the judges will issue a decision that is either unanimous, split, or a majority decision. The dataset includes the amount of victories by each form of decision, as well as by knockout or by the on-hand doctor stopping the fight. Our dataset also provides the current winning or losing streak for each fighter, which can be useful as it can reflect the momentum and confidence a fighter may possess.



The plots above provide a better idea on the breakdown of the amount of male and female fights in UFC's history, as well as the results based on which corner was victorious. Similar to men's and women's lacrosse or men's and women's soccer, we believe that men's and women's UFC fights should be considered different sports given the difference in fighting style. Men's MMA is centered around wrestling, while women's MMA is centered around jiu-jitsu and judo with a strong preference for striking than grappling seen more commonly on the men's side. Since roughly 92% of the data are men's UFC fights, we will remove the observations in which the two fighters are female. The right plot above reveals that about 2/3 of fights with a winner are assigned to the red corner. Since the corners are simply a classifier and we are conerned that our models will be biased towards the red corner, we will randomize the fighters within each fight and reassign their corresponding statistics if necessary.



Another limitation within our data is that for a fighter's first career UFC fight, they have no historical data and thus their respective columns are N/A in our dataset. Since our modeling techniques require clean data without missing data, we will need to remove instances of a fighter's first fight.



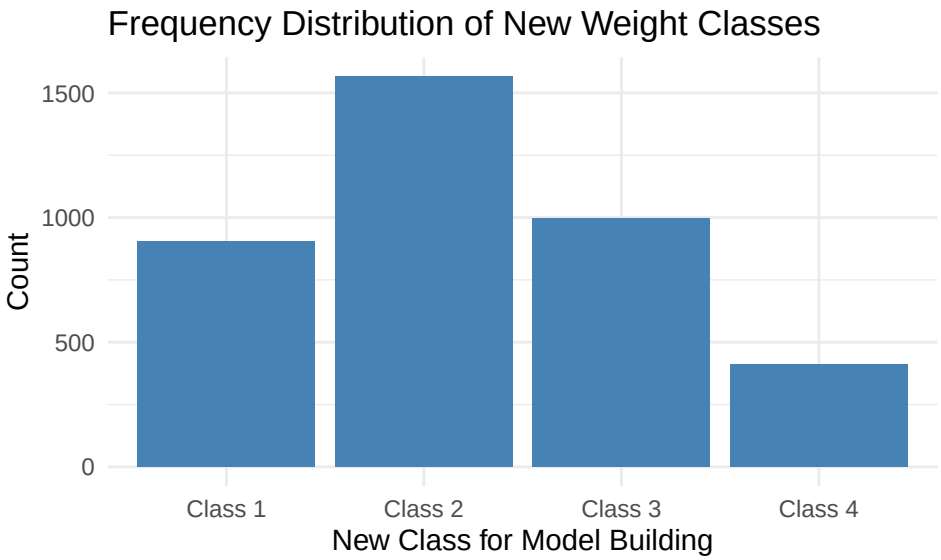
After shuffling the data, we see a much more even distribution of fight winners between the red and blue corner. Similar to comparing the difference between men's and women's MMA, we believe the fighting styles begin to differ as fighters increase in weight class. Therefore, rather than creating nine different models for each class, we will group weight classes together as seen below:

Table 1: Frequency of Weight Class Fights

new_class	weight_class	max_weight	count
Class 1	Flyweight	125	173
Class 1	Bantamweight	135	331
Class 1	Featherweight	145	401
Class 2	Lightweight	155	782
Class 2	Welterweight	170	783
Class 3	Middleweight	185	582
Class 3	LightHeavyweight	205	416
Class 4	Heavyweight	265	381
Class 4	OpenWeight	300	29

Table 2: Avg Attempt of Different Style Attacks

new_class	Rmean_head	Rmean_body	Rmean_leg	Rmean_clinch	Rmean_ground	Rmean_ctrltime
Class 1	73.7	12.0	8.0	7.5	8.5	143.6
Class 2	62.8	10.0	6.9	7.8	8.3	150.5
Class 3	51.9	7.8	6.0	8.0	9.1	138.7
Class 4	43.5	6.3	4.8	6.6	8.9	105.7



The plot above displays the new distribution of observations by the new weight classes. While we were unable to create a completely even split of observations, we believe we have enough fights within each new segmentation to proceed in fitting four models. The table below shows the average attempts of different types of strikes for one classification of the fighters. We see that numbers tend to decrease for standing strikes such as the head, body, and leg as the weight class increases. This could be because heavier fighters prefer to spend more of the fight wrestling on the floor, or heavier fights typically lasting less time, with such cases not mutually exclusive.

Methodology

Before constructing our four models, we first want to remove variables within our dataset that are either redundant or will not provide useful information in predicting a winner. We will remove many of the multi-

class variables, such as the red and blue fighter's names and the referee's name, that would significantly increase the complexity of our models. We will also remove the amount of draws a previous fighter has in their career as all rows equate to zero for both the red and blue fighters. Elastic net model creation requires that the data is clean and does not contain missing values. Although we removed many of the initial missing values for when it was a fighter's UFC debut, there are other cases in our dataset in which a fighter's age or reach is not recorded. We considered imputing the data using averages, but ultimately decided to remove these rows alltogther to avoid potential biases that may arise from imputation. We also believed our sample size was large enough to construct our models.

```
## # A tibble: 1 x 2
##   R_draw count
##    <dbl> <int>
## 1      0  3878
```

```
## # A tibble: 1 x 2
##   B_draw count
##    <dbl> <int>
## 1      0  3878
```

To create the four separate models for the weight class segmentation, we have to create four separate data frames from which we can create training and test splits. Additionally, since the structure of our data set includes the number of attempts and successes for a given strike, as well as current win streak and overall wins on a fighter's record, we suspect that many potential predictors will be extremely multicollinear. To combat this, we plan to use ridge regression to introduce bias that can lower the variance of the estimates. However, since our models contain several predictors and ridge regression does not perform variable selection, we want to use lasso regression as well to assign estimates of irrelevant predictors to zero. The combination of both methods is called elastic net regression a regression model that includes both the L1 penalty of Lasso and the L2 penalty of Ridge regression. Elastic Net combines both L2 and L1 penalties of ridge regression and lasso. It controls the mixing of the two penalties through a parameter (lambda).

```
##                      Var1                    Var2 Correlation
## 1      B_avg_SIG_STR_att          B_avg_HEAD_att   0.9776199
## 2      B_avg_SIG_STR_att      B_avg_DISTANCE_att   0.9697865
## 3  B_avg_opp_SIG_STR_att      B_avg_opp_HEAD_att   0.9821888
## 4  B_avg_opp_SIG_STR_att  B_avg_opp_DISTANCE_att   0.9733084
## 5         B_avg_HEAD_att      B_avg_DISTANCE_att   0.9514359
## 6     B_avg_opp_HEAD_att  B_avg_opp_DISTANCE_att   0.9592315
## 7         B_avg_BODY_att       B_avg_BODY_landed   0.9566996
## 8          B_avg_LEG_att        B_avg_LEG_landed   0.9690741
## 9      B_avg_opp_LEG_att    B_avg_opp_LEG_landed   0.9665108
## 10      B_avg_CLINCH_att     B_avg_CLINCH_landed   0.9641752
## 11  B_avg_opp_CLINCH_att B_avg_opp_CLINCH_landed   0.9643909
## 12      B_avg_GROUND_att     B_avg_GROUND_landed   0.9708085
## 13  B_avg_opp_GROUND_att B_avg_opp_GROUND_landed   0.9734192
## 14     R_avg_SIG_STR_att          R_avg_HEAD_att   0.9769091
## 15     R_avg_SIG_STR_att      R_avg_DISTANCE_att   0.9722080
## 16 R_avg_opp_SIG_STR_att R_avg_opp_TOTAL_STR_att   0.9517534
## 17 R_avg_opp_SIG_STR_att      R_avg_opp_HEAD_att   0.9827113
## 18 R_avg_opp_SIG_STR_att  R_avg_opp_DISTANCE_att   0.9778736
## 19        R_avg_HEAD_att      R_avg_DISTANCE_att   0.9561058
## 20    R_avg_opp_HEAD_att  R_avg_opp_DISTANCE_att   0.9625529
## 21        R_avg_BODY_att       R_avg_BODY_landed   0.9531166
## 22         R_avg_LEG_att        R_avg_LEG_landed   0.9792124
```

```
## 23     R_avg_opp_LEG_att    R_avg_opp_LEG_landed   0.9772406
## 24      R_avg_CLINCH_att     R_avg_CLINCH_landed   0.9694010
## 25  R_avg_opp_CLINCH_att R_avg_opp_CLINCH_landed   0.9598072
## 26      R_avg_GROUND_att     R_avg_GROUND_landed   0.9751036
## 27  R_avg_opp_GROUND_att R_avg_opp_GROUND_landed   0.9706941
```

	Var1	Var2	Correlation
1	B_avg_SIG_STR_att	B_avg_SIG_STR_landed	0.9143492
2	B_avg_SIG_STR_att	B_avg_TOTAL_STR_att	0.9157888
3	B_avg_SIG_STR_att	B_avg_HEAD_att	0.9754681
4	B_avg_SIG_STR_att	B_avg_HEAD_landed	0.8422348
5	B_avg_SIG_STR_att	B_avg_DISTANCE_att	0.9610712
6	B_avg_SIG_STR_att	B_avg_DISTANCE_landed	0.8983882
7	B_avg_SIG_STR_att	B_avg_opp_DISTANCE_att	0.8131356
8	B_avg_SIG_STR_landed	B_avg_TOTAL_STR_att	0.8821178
9	B_avg_SIG_STR_landed	B_avg_HEAD_att	0.8627591
10	B_avg_SIG_STR_landed	B_avg_HEAD_landed	0.9178178
11	B_avg_SIG_STR_landed	B_avg_DISTANCE_att	0.8274141
12	B_avg_SIG_STR_landed	B_avg_DISTANCE_landed	0.8993264
13	B_avg_opp_SIG_STR_att	B_avg_opp_SIG_STR_landed	0.9055164
14	B_avg_opp_SIG_STR_att	B_avg_opp_TOTAL_STR_att	0.9310837
15	B_avg_opp_SIG_STR_att	B_avg_opp_HEAD_att	0.9826437
16	B_avg_opp_SIG_STR_att	B_avg_opp_HEAD_landed	0.8374686
17	B_avg_opp_SIG_STR_att	B_avg_DISTANCE_att	0.8261820
18	B_avg_opp_SIG_STR_att	B_avg_opp_DISTANCE_att	0.9697257
19	B_avg_opp_SIG_STR_att	B_avg_opp_DISTANCE_landed	0.9073489
20	B_avg_opp_SIG_STR_landed	B_avg_opp_TOTAL_STR_att	0.8700991
21	B_avg_opp_SIG_STR_landed	B_avg_opp_HEAD_att	0.8695749
22	B_avg_opp_SIG_STR_landed	B_avg_opp_HEAD_landed	0.9354774
23	B_avg_opp_SIG_STR_landed	B_avg_opp_DISTANCE_att	0.8270996
24	B_avg_opp_SIG_STR_landed	B_avg_opp_DISTANCE_landed	0.9152108
25	B_avg_TOTAL_STR_att	B_avg_TOTAL_STR_landed	0.8890515
26	B_avg_TOTAL_STR_att	B_avg_HEAD_att	0.8896775
27	B_avg_TOTAL_STR_att	B_avg_HEAD_landed	0.8150329
28	B_avg_TOTAL_STR_att	B_avg_DISTANCE_att	0.8145068
29	B_avg_opp_TOTAL_STR_att	B_avg_opp_TOTAL_STR_landed	0.8714003
30	B_avg_opp_TOTAL_STR_att	B_avg_opp_HEAD_att	0.9168587
31	B_avg_opp_TOTAL_STR_att	B_avg_opp_HEAD_landed	0.8134297
32	B_avg_opp_TOTAL_STR_att	B_avg_opp_DISTANCE_att	0.8539701
33	B_avg_HEAD_att	B_avg_HEAD_landed	0.8738385
34	B_avg_HEAD_att	B_avg_DISTANCE_att	0.9371935
35	B_avg_HEAD_att	B_avg_DISTANCE_landed	0.8485865
36	B_avg_HEAD_landed	B_avg_DISTANCE_landed	0.8068470
37	B_avg_opp_HEAD_att	B_avg_opp_HEAD_landed	0.8625363
38	B_avg_opp_HEAD_att	B_avg_DISTANCE_att	0.8051418
39	B_avg_opp_HEAD_att	B_avg_opp_DISTANCE_att	0.9505855
40	B_avg_opp_HEAD_att	B_avg_opp_DISTANCE_landed	0.8695815
41	B_avg_opp_HEAD_landed	B_avg_opp_DISTANCE_landed	0.8301938
42	B_avg_BODY_att	B_avg_BODY_landed	0.9544961
43	B_avg_opp_BODY_att	B_avg_opp_BODY_landed	0.9451332
44	B_avg_LEG_att	B_avg_LEG_landed	0.9806243
45	B_avg_opp_LEG_att	B_avg_opp_LEG_landed	0.9739077
46	B_avg_DISTANCE_att	B_avg_DISTANCE_landed	0.9321247
47	B_avg_DISTANCE_att	B_avg_opp_DISTANCE_att	0.8588292

```
## 48        B_avg_DISTANCE_att  B_avg_opp_DISTANCE_landed   0.8228732
## 49     B_avg_DISTANCE_landed     B_avg_opp_DISTANCE_att   0.8160645
## 50    B_avg_opp_DISTANCE_att  B_avg_opp_DISTANCE_landed   0.9258966
## 51          B_avg_CLINCH_att        B_avg_CLINCH_landed   0.9668683
## 52      B_avg_opp_CLINCH_att    B_avg_opp_CLINCH_landed   0.9598881
## 53          B_avg_GROUND_att        B_avg_GROUND_landed   0.9704378
## 54      B_avg_opp_GROUND_att    B_avg_opp_GROUND_landed   0.9722015
## 55     B_total_rounds_fought                    B_wins   0.9349517
## 56     B_total_rounds_fought                  B_losses   0.8433440
## 57      B_longest_win_streak                    B_wins   0.8267300
## 58         R_avg_SIG_STR_att       R_avg_SIG_STR_landed   0.9243669
## 59         R_avg_SIG_STR_att      R_avg_opp_SIG_STR_att   0.8001668
## 60         R_avg_SIG_STR_att        R_avg_TOTAL_STR_att   0.9284102
## 61         R_avg_SIG_STR_att             R_avg_HEAD_att   0.9801963
## 62         R_avg_SIG_STR_att          R_avg_HEAD_landed   0.8563531
## 63         R_avg_SIG_STR_att         R_avg_DISTANCE_att   0.9719546
## 64         R_avg_SIG_STR_att      R_avg_DISTANCE_landed   0.9151493
## 65         R_avg_SIG_STR_att     R_avg_opp_DISTANCE_att   0.8205181
## 66      R_avg_SIG_STR_landed        R_avg_TOTAL_STR_att   0.8852651
## 67      R_avg_SIG_STR_landed             R_avg_HEAD_att   0.8867078
## 68      R_avg_SIG_STR_landed          R_avg_HEAD_landed   0.9311218
## 69      R_avg_SIG_STR_landed         R_avg_DISTANCE_att   0.8593205
## 70      R_avg_SIG_STR_landed      R_avg_DISTANCE_landed   0.9238372
## 71     R_avg_opp_SIG_STR_att   R_avg_opp_SIG_STR_landed   0.9216880
## 72     R_avg_opp_SIG_STR_att    R_avg_opp_TOTAL_STR_att   0.9339666
## 73     R_avg_opp_SIG_STR_att         R_avg_opp_HEAD_att   0.9814099
## 74     R_avg_opp_SIG_STR_att      R_avg_opp_HEAD_landed   0.8547824
## 75     R_avg_opp_SIG_STR_att         R_avg_DISTANCE_att   0.8207233
## 76     R_avg_opp_SIG_STR_att      R_avg_DISTANCE_landed   0.8104700
## 77     R_avg_opp_SIG_STR_att     R_avg_opp_DISTANCE_att   0.9740427
## 78     R_avg_opp_SIG_STR_att  R_avg_opp_DISTANCE_landed   0.9135800
## 79  R_avg_opp_SIG_STR_landed    R_avg_opp_TOTAL_STR_att   0.8904634
## 80  R_avg_opp_SIG_STR_landed R_avg_opp_TOTAL_STR_landed   0.8030488
## 81  R_avg_opp_SIG_STR_landed         R_avg_opp_HEAD_att   0.8829509
## 82  R_avg_opp_SIG_STR_landed      R_avg_opp_HEAD_landed   0.9361684
## 83  R_avg_opp_SIG_STR_landed     R_avg_opp_DISTANCE_att   0.8636868
## 84  R_avg_opp_SIG_STR_landed  R_avg_opp_DISTANCE_landed   0.9359944
## 85       R_avg_TOTAL_STR_att     R_avg_TOTAL_STR_landed   0.8938762
## 86       R_avg_TOTAL_STR_att             R_avg_HEAD_att   0.9116158
## 87       R_avg_TOTAL_STR_att          R_avg_HEAD_landed   0.8293939
## 88       R_avg_TOTAL_STR_att         R_avg_DISTANCE_att   0.8552787
## 89       R_avg_TOTAL_STR_att      R_avg_DISTANCE_landed   0.8003391
## 90   R_avg_opp_TOTAL_STR_att R_avg_opp_TOTAL_STR_landed   0.8802815
## 91   R_avg_opp_TOTAL_STR_att         R_avg_opp_HEAD_att   0.9148592
## 92   R_avg_opp_TOTAL_STR_att      R_avg_opp_HEAD_landed   0.8289642
## 93   R_avg_opp_TOTAL_STR_att     R_avg_opp_DISTANCE_att   0.8620021
## 94   R_avg_opp_TOTAL_STR_att  R_avg_opp_DISTANCE_landed   0.8089473
## 95            R_avg_HEAD_att          R_avg_HEAD_landed   0.8871980
## 96            R_avg_HEAD_att         R_avg_DISTANCE_att   0.9509603
## 97            R_avg_HEAD_att      R_avg_DISTANCE_landed   0.8764212
## 98            R_avg_HEAD_att     R_avg_opp_DISTANCE_att   0.8001721
## 99         R_avg_HEAD_landed      R_avg_DISTANCE_landed   0.8416924
## 100       R_avg_opp_HEAD_att      R_avg_opp_HEAD_landed   0.8783261
## 101       R_avg_opp_HEAD_att         R_avg_DISTANCE_att   0.8079495
```

```
## 102       R_avg_opp_HEAD_att     R_avg_opp_DISTANCE_att   0.9570363
## 103       R_avg_opp_HEAD_att  R_avg_opp_DISTANCE_landed   0.8783584
## 104    R_avg_opp_HEAD_landed  R_avg_opp_DISTANCE_landed   0.8671617
## 105           R_avg_BODY_att          R_avg_BODY_landed   0.9543110
## 106       R_avg_opp_BODY_att      R_avg_opp_BODY_landed   0.9542727
## 107            R_avg_LEG_att           R_avg_LEG_landed   0.9802785
## 108        R_avg_opp_LEG_att       R_avg_opp_LEG_landed   0.9808604
## 109       R_avg_DISTANCE_att      R_avg_DISTANCE_landed   0.9388924
## 110       R_avg_DISTANCE_att     R_avg_opp_DISTANCE_att   0.8504474
## 111       R_avg_DISTANCE_att  R_avg_opp_DISTANCE_landed   0.8109810
## 112    R_avg_DISTANCE_landed     R_avg_opp_DISTANCE_att   0.8442854
## 113    R_avg_DISTANCE_landed  R_avg_opp_DISTANCE_landed   0.8202149
## 114   R_avg_opp_DISTANCE_att  R_avg_opp_DISTANCE_landed   0.9335455
## 115         R_avg_CLINCH_att        R_avg_CLINCH_landed   0.9663859
## 116     R_avg_opp_CLINCH_att    R_avg_opp_CLINCH_landed   0.9693263
## 117         R_avg_GROUND_att        R_avg_GROUND_landed   0.9662308
## 118     R_avg_opp_GROUND_att    R_avg_opp_GROUND_landed   0.9580675
## 119    R_total_rounds_fought                    R_wins   0.9409881
## 120    R_total_rounds_fought                  R_losses   0.8650970
## 121     R_longest_win_streak                    R_wins   0.8196516
```

In looking at the correlation coefficients between predictors for each class, we notice that coefficients for the same type of strike exceed 0.95 (such has avg strikes attempted and landed to the opponent's head). Correlation coefficients are also extremely high for significant strike predictors and other forms of strikes, such as head, body, and leg. Therefore, once we fit each of the models, we will also construct anova tables to determine if the inclusion of interactions amongst the selected variables help improve the model.

Since we plan to use elastic net that contain, we standardized the numerical predictors in our data. In doing so, we first created our training and test splits. From there, we scaled each split using the mean and standard deviation of the training data so there would not be any information leakage into our test set as we need to consider it as establishing new data. Elastic net also does not allow for multi-class categorical variables, so we need to transform the categorical variables (including our response variable) into numerical format so they can be used in Elastic Net. We can do this by using dummy variables. We will not include a dummy variable indicating a Winner for the blue corner so our models are only in the context of Red winning or losing (1 = Red win, 0 = Red lose). Using a 70/30 training and test split, we can use cross validation for each alpha value to find the optimal lambda. Alpha is the mixing parameter between Lasso (alpha = 1) and Ridge (alpha = 0) regression. We will find the combination that gives the best performance (e.g., the lowest deviance), fit the elastic net model using the optimal alpha and lambda values, and finally evaluate the model on test set to check its performance.

Accuracy: proportion of the total number of predictions that were correct. Precision: ratio of correctly predicted positive observations to the total predicted positives. Recall (Sensitivity): ratio of correctly predicted positive observations to all observations in the actual class. F1-Score: weighted average of Precision and Recall.

Class 1 Model

```
## Analysis of Deviance Table
##
## Model 1: WinnerRed ~ B_avg_KD + B_avg_opp_SIG_STR_pct + B_avg_TD_att +
##     B_avg_opp_TD_att + B_avg_opp_HEAD_landed + B_avg_CLINCH_att +
##     B_win_by_Decision_Majority + B_win_by_Decision_Unanimous +
##     B_win_by_Submission + R_avg_KD + R_avg_opp_SIG_STR_pct +
##     R_avg_opp_TD_pct + R_avg_opp_SUB_ATT + R_avg_REV + R_avg_SIG_STR_landed +
```

```
##     R_avg_BODY_landed + R_avg_LEG_att + R_avg_opp_CLINCH_att +
##     R_win_by_KO.TKO + R_current_win_streak + B_age + R_age +
##     B_StanceSwitch
## Model 2: WinnerRed ~ B_avg_KD + B_avg_opp_SIG_STR_pct + B_avg_TD_att +
##     B_avg_opp_TD_att + B_avg_opp_HEAD_landed + B_avg_CLINCH_att +
##     B_win_by_Decision_Majority + B_win_by_Decision_Unanimous +
##     B_win_by_Submission + R_avg_KD + R_avg_opp_SIG_STR_pct +
##     R_avg_opp_TD_pct + R_avg_opp_SUB_ATT + R_avg_REV + R_avg_SIG_STR_landed +
##     R_avg_BODY_landed + R_avg_LEG_att + R_avg_opp_CLINCH_att +
##     R_current_win_streak + R_win_by_KO.TKO + B_age + R_age +
##     B_StanceSwitch + R_avg_REV * R_avg_SIG_STR_landed + R_avg_BODY_landed *
##     R_avg_SIG_STR_landed + R_avg_opp_TD_pct * R_avg_opp_SIG_STR_pct
##   Resid. Df Resid. Dev Df Deviance
## 1       577     711.12
## 2       574     709.79  3   1.3303
```

```
##
## Call:
## glm(formula = WinnerRed ~ B_avg_KD + B_avg_opp_SIG_STR_pct +
##     B_avg_TD_att + B_avg_opp_TD_att + B_avg_opp_HEAD_landed +
##     B_avg_CLINCH_att + B_win_by_Decision_Majority + B_win_by_Decision_Unanimous +
##     B_win_by_Submission + R_avg_KD + R_avg_opp_SIG_STR_pct +
##     R_avg_opp_TD_pct + R_avg_opp_SUB_ATT + R_avg_REV + R_avg_SIG_STR_landed +
##     R_avg_BODY_landed + R_avg_LEG_att + R_avg_opp_CLINCH_att +
##     R_win_by_KO.TKO + R_current_win_streak + B_age + R_age +
##     B_StanceSwitch, family = "binomial", data = class1train_scaled)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1268  -0.9922  -0.4079   0.9910   2.2656
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    0.06228    4.36093   0.014 0.988606
## B_avg_KD                       0.14680    0.10025   1.464 0.143109
## B_avg_opp_SIG_STR_pct          0.25139    0.09632   2.610 0.009053 **
## B_avg_TD_att                  -0.11499    0.10561  -1.089 0.276248
## B_avg_opp_TD_att              -0.12019    0.10406  -1.155 0.248090
## B_avg_opp_HEAD_landed          0.15568    0.10142   1.535 0.124784
## B_avg_CLINCH_att              -0.16117    0.11603  -1.389 0.164797
## B_win_by_Decision_Majority     1.32965   53.30893   0.025 0.980101
## B_win_by_Decision_Unanimous   -0.43715    0.11177  -3.911 9.19e-05 ***
## B_win_by_Submission           -0.12775    0.09526  -1.341 0.179872
## R_avg_KD                       0.05817    0.10086   0.577 0.564131
## R_avg_opp_SIG_STR_pct         -0.26639    0.10252  -2.598 0.009368 **
## R_avg_opp_TD_pct              -0.10338    0.09919  -1.042 0.297295
## R_avg_opp_SUB_ATT             -0.14436    0.10501  -1.375 0.169225
## R_avg_REV                     -0.11704    0.09906  -1.182 0.237389
## R_avg_SIG_STR_landed           0.09425    0.15847   0.595 0.551996
## R_avg_BODY_landed              0.24719    0.14003   1.765 0.077531 .
## R_avg_LEG_att                  0.07144    0.11287   0.633 0.526753
## R_avg_opp_CLINCH_att          -0.19665    0.10221  -1.924 0.054352 .
## R_win_by_KO.TKO                0.13943    0.10947   1.274 0.202771
## R_current_win_streak           0.24652    0.10112   2.438 0.014775 *
```

```
## B_age                        0.35136    0.10410   3.375 0.000737 ***
## R_age                       -0.40040    0.09896  -4.046 5.20e-05 ***
## B_StanceSwitch              -0.84015    0.45164  -1.860 0.062853 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 832.79  on 600  degrees of freedom
## Residual deviance: 711.12  on 577  degrees of freedom
## AIC: 759.12
##
## Number of Fisher Scoring iterations: 14
```

For class 1 our optimal alpha value is 0.75, which results in the variable selection of 22 predictors. Given the high coefficient values we found earlier in our methodology, we attempted to include three interaction terms. In a new model with the selected variables, we added the interaction of the average amount of reversal strikes landed for the red fighter and the average amount of significant strikes landed by the red fighter, the interaction of the average amount of body strikes landed and the average amount of significant strikes landed by the Red Fighter, and the average amount of takedowns and the average amount of significant strikes previous opponents have landed on the Red fighter. Additionally, we also included a quadratic effect for the average amount of significant strikes landed under the assumption that as a fighter delivers more damage to their opponent the probability of winning would increase exponentially. However, we failed to reject the null hypothesis in favor for the base model when each if the added terms were included and the anova test was run.

```
##                   B_avg_KD       B_avg_opp_SIG_STR_pct
##                   1.210199                    1.153048
##                B_avg_TD_att            B_avg_opp_TD_att
##                   1.367588                    1.339988
##       B_avg_opp_HEAD_landed            B_avg_CLINCH_att
##                   1.208732                    1.363585
##  B_win_by_Decision_Majority B_win_by_Decision_Unanimous
##                   1.000000                    1.474890
##         B_win_by_Submission                   R_avg_KD
##                   1.123130                    1.220226
##       R_avg_opp_SIG_STR_pct            R_avg_opp_TD_pct
##                   1.154675                    1.192042
##           R_avg_opp_SUB_ATT                  R_avg_REV
##                   1.184695                    1.158500
##        R_avg_SIG_STR_landed           R_avg_BODY_landed
##                   2.945254                    2.363794
##                R_avg_LEG_att        R_avg_opp_CLINCH_att
##                   1.538633                    1.278402
##             R_win_by_KO.TKO        R_current_win_streak
##                   1.409574                    1.175030
##                      B_age                       R_age
##                   1.282910                    1.145902
##              B_StanceSwitch
##                   1.096930
```

Since our elastic net model for class 1 recieved a 0.75 alpha value, we hypothesized that some of our predictors were very multicollinear. However, in running a VIF test on our chosen predictors we do not spot any that

would signify high correlation with each other which would justify either removing the predictor or testing further interactions.

Table 3: Class 1 Model Stats

	Values
Accuracy	0.531
Precision	0.504
Recall	0.516
F1	0.510

```
## [1] 12
```

Table 4: Class 1 Model Stats Simple

	Values
Accuracy	0.581
Precision	0.559
Recall	0.541
F1	0.550

Since the use of interactions and the vif did not change our model, we decided to look at the statistically significant predictors and evaluate the model's performance with the subset of predictors. We discovered that our model performs marginally better with less predictors, and given that our goal is to maximize predictive accuracy while also being able to notify trainers the most important statistics to help improve their fighters, we will proceed with the condensed model.

$$log(P(Winner = Red)/1-P(Winner = Red)) = B_0+B_1*B\_avg\_opp\_SIG\_STR\_pct+B_2*B\_win\_by\_Decision\_Unani$$

$$B_3 * R\_avg\_opp\_SIG\_STR\_pct + B_4 * R\_avg\_BODY\_landed+$$

$$B_5 * R\_avg\_opp\_CLINCH\_att + B_6 * R\_current\_win\_streak+$$

$$B_7 * B\_age + B_8 * R\_age + B_9 * BstanceSwitch$$

Class 2 Model

```
## 142 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                 -0.029416655
## B_avg_KD                     .
## B_avg_opp_KD                 .
## B_avg_SIG_STR_pct            .
## B_avg_opp_SIG_STR_pct        0.099518667
## B_avg_TD_pct                 .
## B_avg_opp_TD_pct             .
## B_avg_SUB_ATT                .
## B_avg_opp_SUB_ATT           -0.029356834
## B_avg_REV                    0.033409578
## B_avg_opp_REV                .
## B_avg_SIG_STR_att            .
```

```
## B_avg_SIG_STR_landed          .
## B_avg_opp_SIG_STR_att         .
## B_avg_opp_SIG_STR_landed      .
## B_avg_TOTAL_STR_att           .
## B_avg_TOTAL_STR_landed        .
## B_avg_opp_TOTAL_STR_att       .
## B_avg_opp_TOTAL_STR_landed    .
## B_avg_TD_att                 -0.063278097
## B_avg_TD_landed              -0.123488979
## B_avg_opp_TD_att              .
## B_avg_opp_TD_landed           .
## B_avg_HEAD_att                .
## B_avg_HEAD_landed             .
## B_avg_opp_HEAD_att            .
## B_avg_opp_HEAD_landed         .
## B_avg_BODY_att                .
## B_avg_BODY_landed             .
## B_avg_opp_BODY_att            .
## B_avg_opp_BODY_landed         .
## B_avg_LEG_att                 .
## B_avg_LEG_landed              .
## B_avg_opp_LEG_att             .
## B_avg_opp_LEG_landed          .
## B_avg_DISTANCE_att            .
## B_avg_DISTANCE_landed         .
## B_avg_opp_DISTANCE_att        .
## B_avg_opp_DISTANCE_landed     .
## B_avg_CLINCH_att              .
## B_avg_CLINCH_landed           0.002640710
## B_avg_opp_CLINCH_att          .
## B_avg_opp_CLINCH_landed       .
## B_avg_GROUND_att              .
## B_avg_GROUND_landed           .
## B_avg_opp_GROUND_att          .
## B_avg_opp_GROUND_landed       .
## B_avg_CTRL_time.seconds.     -0.009778574
## B_avg_opp_CTRL_time.seconds.  .
## B_total_time_fought.seconds. -0.004787831
## B_total_rounds_fought         .
## B_total_title_bouts          -0.062200357
## B_current_win_streak          .
## B_current_lose_streak         0.030308304
## B_longest_win_streak         -0.004740481
## B_wins                        .
## B_losses                      .
## B_win_by_Decision_Majority    .
## B_win_by_Decision_Split       .
## B_win_by_Decision_Unanimous   .
## B_win_by_KO.TKO               .
## B_win_by_Submission          -0.080852351
## B_win_by_TKO_Doctor_Stoppage  0.026136738
## B_Height_cms                  .
## B_Reach_cms                   .
## B_Weight_lbs                  .
```

```
## R_avg_KD                        .
## R_avg_opp_KD                    .
## R_avg_SIG_STR_pct               0.002287377
## R_avg_opp_SIG_STR_pct           .
## R_avg_TD_pct                    .
## R_avg_opp_TD_pct                .
## R_avg_SUB_ATT                   0.033690100
## R_avg_opp_SUB_ATT               .
## R_avg_REV                       .
## R_avg_opp_REV                   .
## R_avg_SIG_STR_att               .
## R_avg_SIG_STR_landed            .
## R_avg_opp_SIG_STR_att           .
## R_avg_opp_SIG_STR_landed        .
## R_avg_TOTAL_STR_att             .
## R_avg_TOTAL_STR_landed          .
## R_avg_opp_TOTAL_STR_att         .
## R_avg_opp_TOTAL_STR_landed      .
## R_avg_TD_att                    .
## R_avg_TD_landed                 .
## R_avg_opp_TD_att                .
## R_avg_opp_TD_landed             .
## R_avg_HEAD_att                  .
## R_avg_HEAD_landed               0.026743387
## R_avg_opp_HEAD_att              .
## R_avg_opp_HEAD_landed           .
## R_avg_BODY_att                  .
## R_avg_BODY_landed               .
## R_avg_opp_BODY_att              .
## R_avg_opp_BODY_landed           .
## R_avg_LEG_att                   .
## R_avg_LEG_landed                .
## R_avg_opp_LEG_att               .
## R_avg_opp_LEG_landed            .
## R_avg_DISTANCE_att              .
## R_avg_DISTANCE_landed           .
## R_avg_opp_DISTANCE_att          .
## R_avg_opp_DISTANCE_landed       .
## R_avg_CLINCH_att                .
## R_avg_CLINCH_landed             .
## R_avg_opp_CLINCH_att           -0.098570681
## R_avg_opp_CLINCH_landed         .
## R_avg_GROUND_att                0.007266924
## R_avg_GROUND_landed             .
## R_avg_opp_GROUND_att            .
## R_avg_opp_GROUND_landed         .
## R_avg_CTRL_time.seconds.        .
## R_avg_opp_CTRL_time.seconds. -0.098112340
## R_total_time_fought.seconds.    .
## R_total_rounds_fought           .
## R_total_title_bouts             .
## R_current_win_streak            0.061261605
## R_current_lose_streak           .
## R_longest_win_streak            0.016685470
```

```
## R_wins                         .
## R_losses                       .
## R_win_by_Decision_Majority     .
## R_win_by_Decision_Split       -0.173989902
## R_win_by_Decision_Unanimous    0.043655458
## R_win_by_KO.TKO                .
## R_win_by_Submission            0.004506602
## R_win_by_TKO_Doctor_Stoppage   .
## R_Height_cms                   .
## R_Reach_cms                    .
## R_Weight_lbs                   .
## B_age                          0.281585828
## R_age                         -0.138988895
## id                             .
## title_boutFALSE                .
## title_boutTRUE                 .
## B_StanceOrthodox               0.051470245
## B_StanceSouthpaw              -0.156447680
## B_StanceSwitch                 .
## R_StanceOrthodox               .
## R_StanceSouthpaw               .
## R_StanceSwitch                 .

## Analysis of Variance Table
##
## Model 1: WinnerRed ~ B_avg_opp_SIG_STR_pct + B_avg_opp_SUB_ATT + B_avg_REV +
##     B_avg_TD_att + B_avg_TD_landed + B_avg_CLINCH_landed + B_avg_CTRL_time.seconds. +
##     B_total_time_fought.seconds. + B_total_title_bouts + B_current_lose_streak +
##     B_longest_win_streak + B_win_by_Submission + B_win_by_TKO_Doctor_Stoppage +
##     R_avg_SIG_STR_pct + R_avg_SUB_ATT + R_avg_HEAD_landed + R_avg_opp_CLINCH_att +
##     R_avg_GROUND_att + R_avg_opp_CTRL_time.seconds. + R_current_win_streak +
##     R_longest_win_streak + R_win_by_Decision_Split + R_win_by_Decision_Unanimous +
##     R_win_by_Submission + B_age + R_age + B_StanceOrthodox +
##     B_StanceSouthpaw
## Model 2: WinnerRed ~ B_avg_opp_SIG_STR_pct + B_avg_opp_SUB_ATT + B_avg_REV +
##     B_avg_TD_att + B_avg_TD_landed + B_avg_CLINCH_landed + B_avg_CTRL_time.seconds. +
##     B_total_time_fought.seconds. + B_total_title_bouts + B_current_lose_streak +
##     B_longest_win_streak + B_win_by_Submission + B_win_by_TKO_Doctor_Stoppage +
##     R_avg_SIG_STR_pct + R_avg_SUB_ATT + poly(R_avg_HEAD_landed,
##     2) + R_avg_opp_CLINCH_att + R_avg_GROUND_att + R_avg_opp_CTRL_time.seconds. +
##     R_current_win_streak + R_longest_win_streak + R_win_by_Decision_Split +
##     R_win_by_Decision_Unanimous + R_win_by_Submission + B_age +
##     R_age + B_StanceOrthodox + B_StanceSouthpaw
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    972 218.31
## 2    971 216.43  1    1.8774 8.4228 0.003789 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test explanation for class 2 models

Table 5: Class 2 Model Stats

	Values
Accuracy	0.606
Precision	0.617
Recall	0.592
F1	0.604

Table 6: Most Important Predictors for Class 2 Red Fighter

	s1		s1
B_age	0.282	R_win_by_Decision_Split	-0.174
B_avg_opp_SIG_STR_pct	0.100	B_StanceSouthpaw	-0.156
R_current_win_streak	0.061	R_age	-0.139
B_StanceOrthodox	0.051	B_avg_TD_landed	-0.123
R_win_by_Decision_Unanimous	0.044	R_avg_opp_CLINCH_att	-0.099
R_avg_SUB_ATT	0.034	R_avg_opp_CTRL_time.seconds.	-0.098
B_avg_REV	0.033	B_win_by_Submission	-0.081
B_current_lose_streak	0.030	B_avg_TD_att	-0.063

Class 3 Model

For class 3 the initial alpha value was zero, meaning the model was solely performing ridge regression and not removing any of our variables.

[!h]

Table 7: Class 3 Model Stats

	Values
Accuracy	0.597
Precision	0.596
Recall	0.697
F1	0.643

Table 8: Most Important Predictors for Class 3 Red Fighter

	s1		s1
B_avg_opp_TOTAL_STR_landed	0.245	R_StanceSwitch	-0.668
R_avg_KD	0.116	R_age	-0.297
B_avg_opp_TD_pct	0.110	B_Weight_lbs	-0.183
R_avg_TD_landed	0.106	B_StanceOpen Stance	-0.178
R_longest_win_streak	0.095	B_avg_opp_TD_att	-0.153
B_StanceOrthodox	0.091	B_avg_TD_att	-0.135
R_Height_cms	0.081	B_avg_opp_LEG_att	-0.134
B_avg_opp_GROUND_landed	0.076	(Intercept)	-0.131

test explanation for class 3 models

Class 4 Model

test explanation for class 4 models

Appendix

Data Dictionary

R_ and B_: Prefix signifies red and blue corner fighter stats respectively

opp: Containing in columns is the action done by the opponent on the fighter

fighter: Name of the fighter

Referee: Referee/On-Hand Doctor of the fight. They are responsible for ending a fight if they believe a fighter is unable to continue

date: Date of the fight

location: Location of the fight

Winner: The corner of the winning fighter. We will turn this into a dummy variable and will serve as our model's response variable

title_bout: True/False indicator of a championship fight for a weightclass

weight_class: Categorical variable indicating the division of the fight. There are nine male divisions and four female divisons. We will reassign the male divisons to create our models

KD: the number of knockdowns

SIG_STR: the of significant strikes 'landed of attempted'

SIG_STR_pct: significant strikes percentage

TOTAL_STR: total strikes 'landed of attempted'

TD: number of takedowns

TD_pct: takedown percentages

SUB_ATT: number of submission attempts

REV: number of reversals landed

HEAD: number significant strikes to the head (att = attempted, landed = successful attempts)

BODY: number of significant strikes to the body (att = attempted, landed = successful attempts)

LEG: number of significant strikes to the leg (att = attempted, landed = successful attempts)

CLINCH: number of significant strikes in the clinch, also known as close quarters (att = attempted, landed = successful attempts)

GROUND: number of significant strikes on the ground (att = attempted, landed = successful attempts)

Stance: the stance of the fighter (orthodox, southpaw, etc.)

Height_cms: the height of the fighter in centimeters

Reach_cms: the reach of the fighter (arm span) in centimeters

Weight_lbs: the weight of the fighter in pounds (lbs)

age: the age of the fighter

current_lose_streak: the amount of consecutive previous fights the fighter has lost (0 if they won their previous fight)

current_win_streak: the amount of consecutive previous fights the fighter has won (0 if they lost their previous fight)

draw: the number of draws in the fighter's ufc career

wins: the number of wins in the fighter's ufc career

losses: the number of losses in the fighter's ufc career

total_rounds_fought: the average of total rounds fought by the fighter

total_time_fought(seconds): the count of total time spent fighting in seconds

total_title_bouts: the total number of title bouts taken part in by the fighter

win_by_Decision_Majority: the number of wins by majority judges decision in the fighter's ufc career (often 2-0 with one judge deciding a draw)

win_by_Decision_Split: the number of wins by split judges decision in the fighter's ufc career (often 2-1 in favor of one fighter)

win_by_Decision_Unanimous: the number of wins by unanimous judges decision in the fighter's ufc career

win_by_KO/TKO: the number of wins by knockout in the fighter's ufc career

win_by_Submission: the number of wins by submission in the fighter's ufc career

win_by_TKO_Doctor_Stoppage: the number of wins by doctor stoppage in the fighter's ufc career