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.

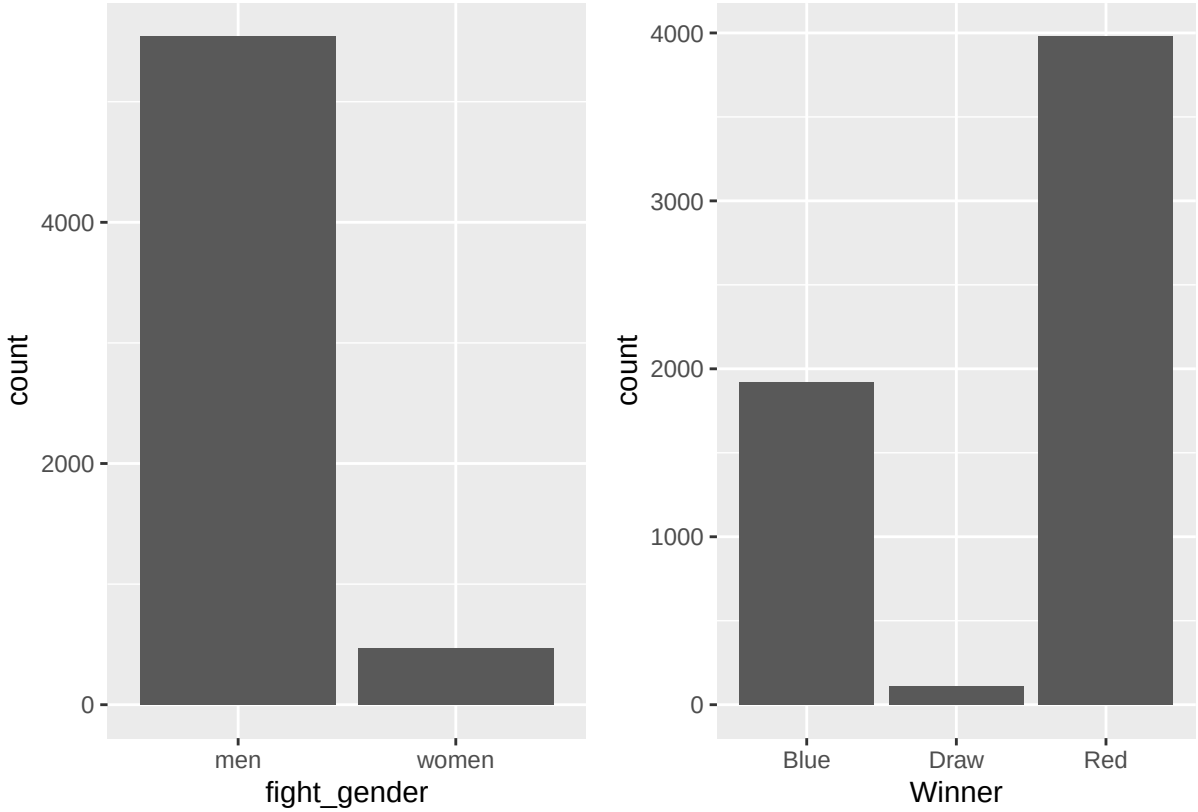
Our data is each UFC fight from 1994-2021, containing each fighter's names, physical information (age, height, weight), the amount of wins and losses in their UFC history, and various fighting data. The fighting data includes the average amount of attempted and landed attacks over their UFC career, as 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

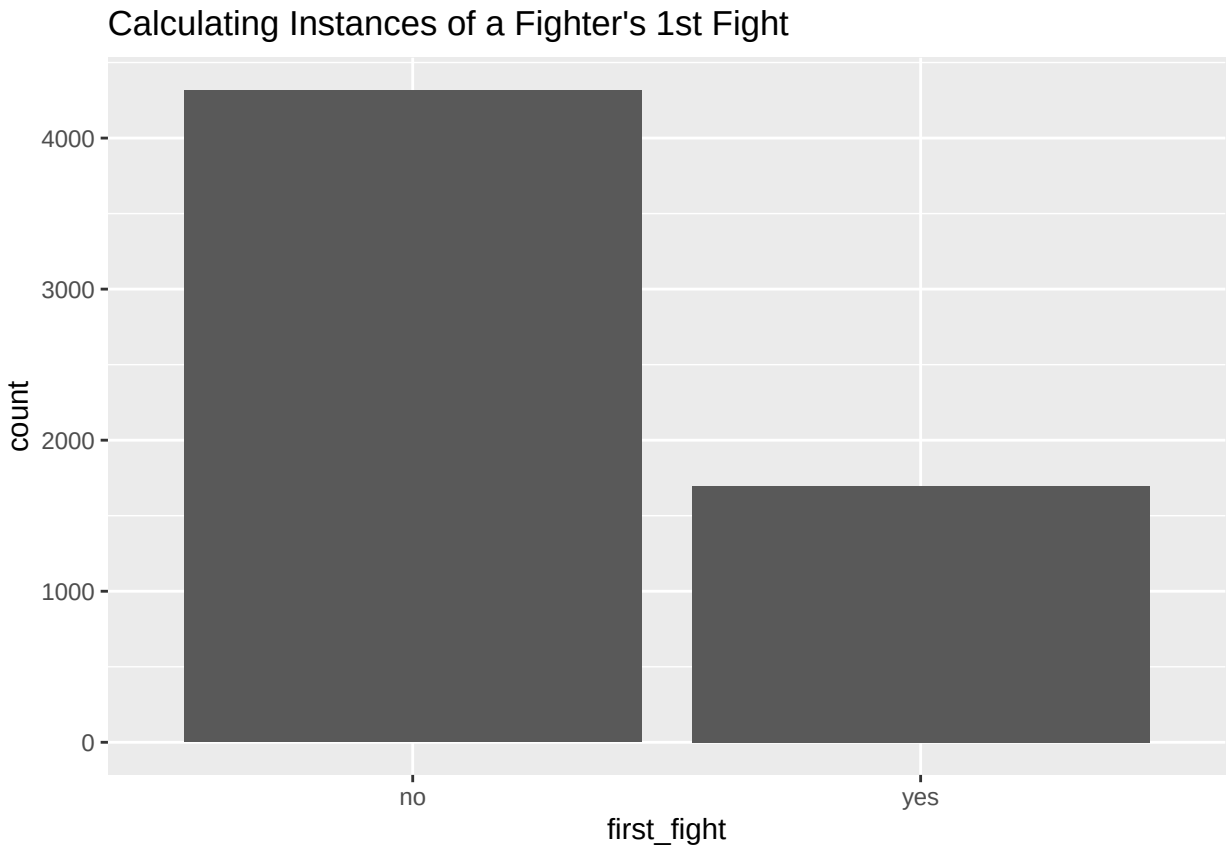
The dataset used for modeling was initially sourced from ufcstats.com, where the data was processed and published on kaggle.com. The dataset contains roughly 6,012 unique fights over a 27 year span, with each

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.



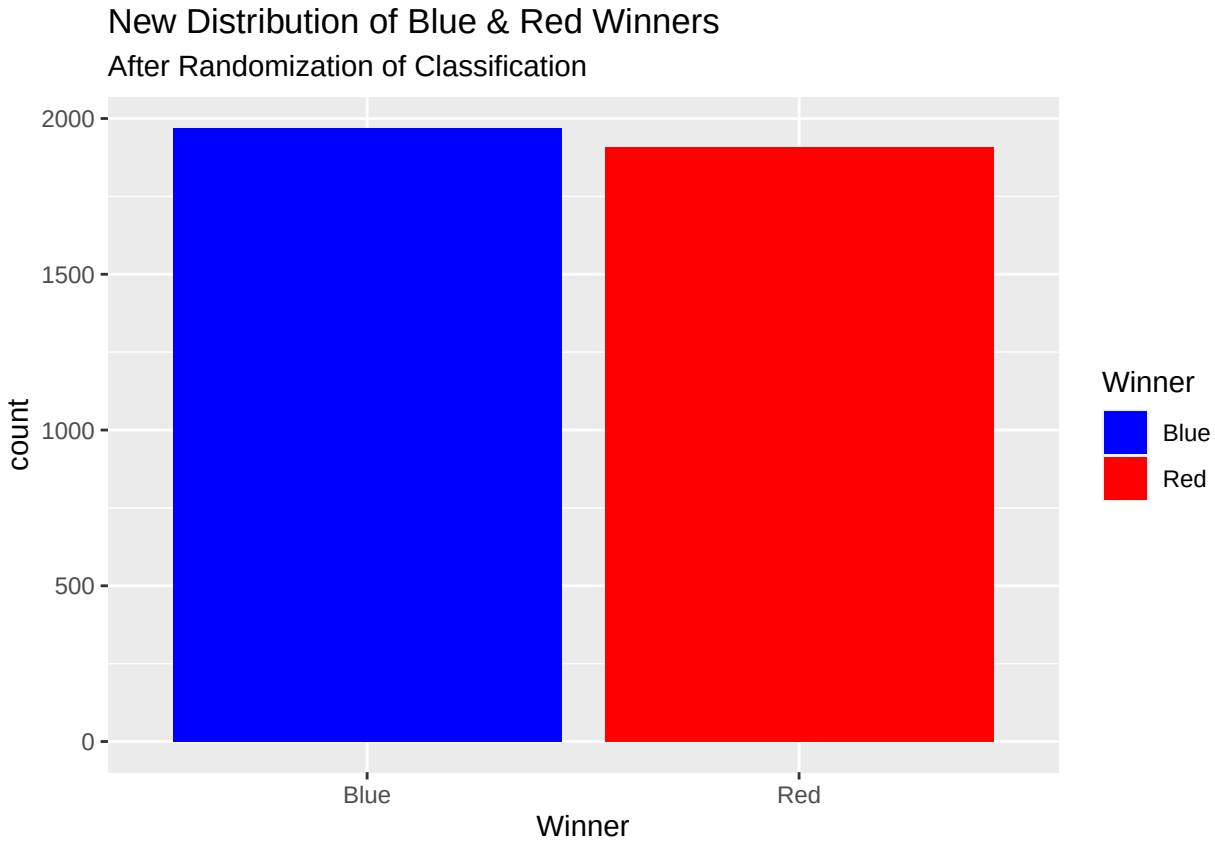
The plots above provide a better idea on the breakdown of the amount of male and female fights in UFC's history, as well as the results based on which corner was victorious. Similar to men's and women's lacrosse or men's and women's soccer, we believe that men's and women's UFC fights should be considered 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.

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



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 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
```

```
## [1] 3878
```

```
## [1] 3511
```

```
## [1] 367
```

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).

Since we plan to use elastic net that contain, we standardized the numerical predictors in our 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 to find the optimal lambda value to penalize our predictors and use lasso to perform variable selection.

Class 1 Model

Table 3: Class 1 Model Stats

	Values
Accuracy	0.539
Precision	0.512
Recall	0.525
F1	0.518

Now moving to models, we can perform cross-validation to find best lambda value. Will use to find the optimal lambda and alpha parameters. Alpha is the mixing parameter between Lasso (alpha = 1) and Ridge

Table 4: Most Important Predictors for Class 1 Red Fighter

	s1		s1
B_age	0.216	B_StanceSwitch	-0.322
R_current_win_streak	0.141	R_age	-0.239
B_avg_opp_SIG_STR_pct	0.137	B_win_by_Decision_Unanimous	-0.186
B_win_by_Decision_Majority	0.136	R_avg_opp_SIG_STR_pct	-0.183
R_avg_BODY_landed	0.120	B_avg_CLINCH_att	-0.089
R_avg_SIG_STR_landed	0.098	B_avg_TD_att	-0.085
R_win_by_KO.TKO	0.077	R_avg_opp_CLINCH_att	-0.074
B_avg_KD	0.053	R_avg_opp_TD_pct	-0.062

(alpha = 0) regression. Run cross-validation for each alpha value to find the optimal lambda. 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. 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 2 Model

Table 5: Class 2 Model Stats

	Values
Accuracy	0.597
Precision	0.610
Recall	0.573
F1	0.591

Table 6: Most Important Predictors for Class 2 Red Fighter

	s1		s1
B_age	0.317	B_StanceSouthpaw	-0.199
B_avg_opp_SIG_STR_pct	0.114	R_win_by_Decision_Split	-0.196
R_current_win_streak	0.064	R_age	-0.164
R_win_by_Decision_Unanimous	0.062	B_avg_TD_landed	-0.133
B_avg_REV	0.051	R_avg_opp_CTRL_time.seconds.	-0.118
B_win_by_TKO_Doctor_Stoppage	0.046	R_avg_opp_CLINCH_att	-0.108
R_avg_SUB_ATT	0.045	B_win_by_Submission	-0.096
B_current_lose_streak	0.040	B_total_title_bouts	-0.080

Test explanation for class 2 models

Class 3 Model

test explanation for class 3 models

[!h]

Table 7: Class 3 Model Stats

	Values
Accuracy	0.626
Precision	0.606
Recall	0.803
F1	0.691

Table 8: Most Important Predictors for Class 3 Red Fighter

	s1		s1
R_StanceOpen Stance	0.088	R_StanceSwitch	-0.320
B_StanceOrthodox	0.056	B_StanceOpen Stance	-0.319
R_StanceSouthpaw	0.055	(Intercept)	-0.178
B_avg_opp_TD_pct	0.053	R_age	-0.101
R_avg_KD	0.049	B_StanceSwitch	-0.098
B_avg_opp_TOTAL_STR_landed	0.048	B_Weight_lbs	-0.069
R_avg_opp_TD_att	0.044	B_Reach_cms	-0.049
B_avg_opp_SIG_STR_pct	0.042	R_losses	-0.049

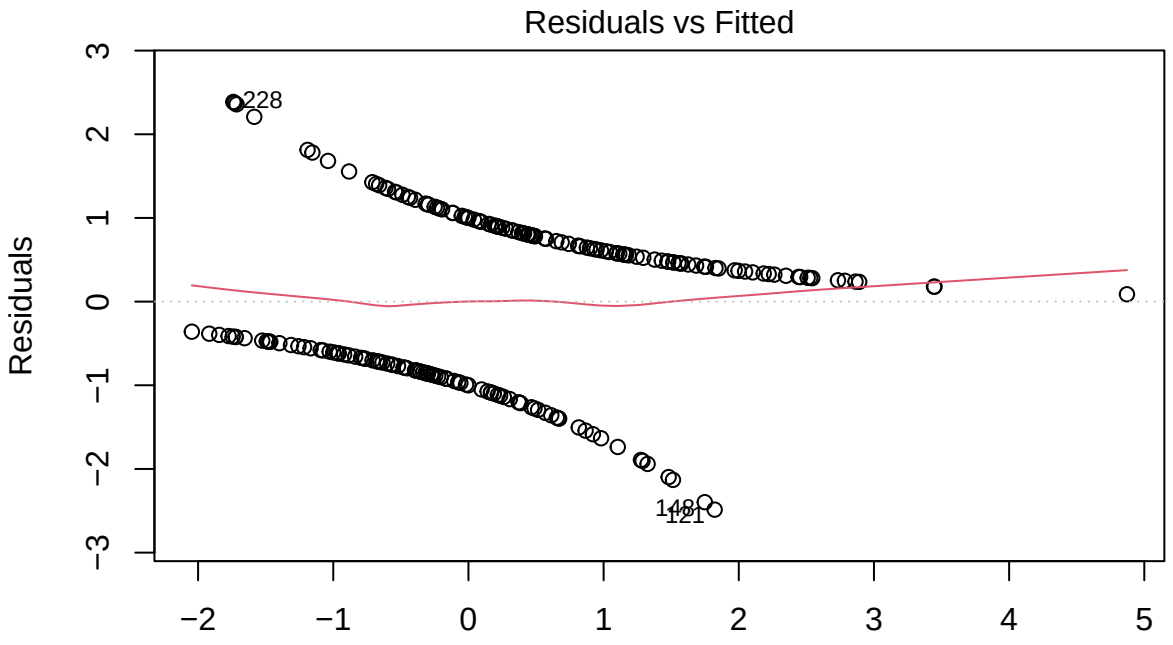
Class 4 Model

Table 9: Class 4 Model Stats

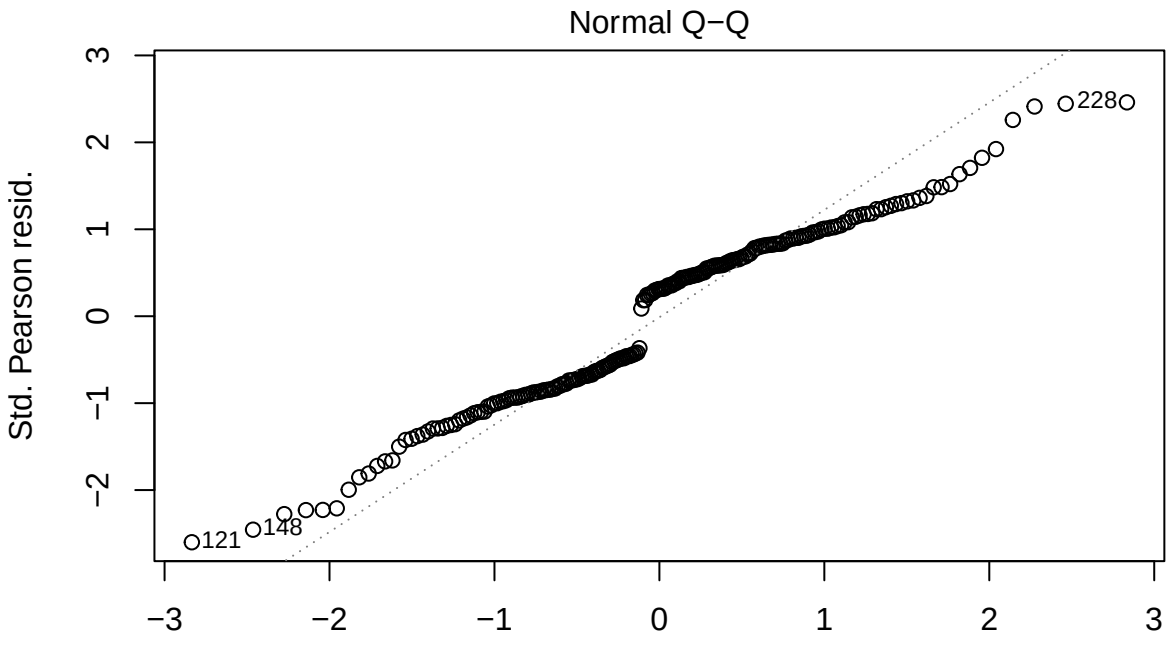
	Values
Accuracy	0.570
Precision	0.645
Recall	0.408
F1	0.500

Table 10: Most Important Predictors for Class 4 Red Fighter

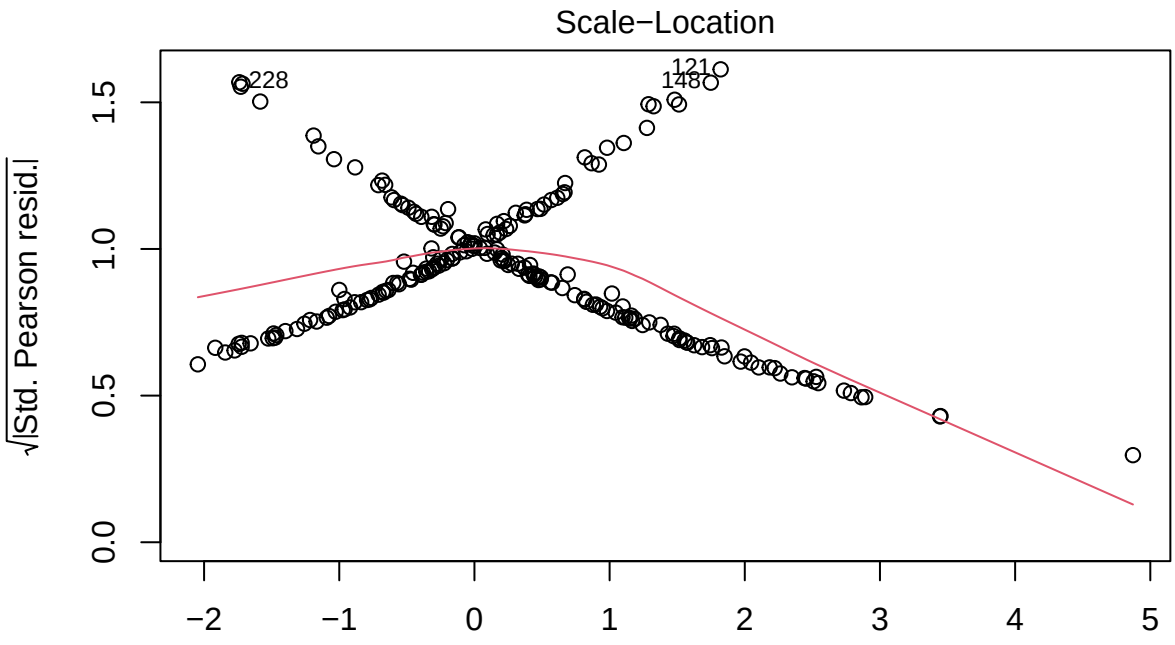
	s1		s1
(Intercept)	0.213	B_Reach_cms	-0.136
B_avg_opp_SIG_STR_pct	0.138	R_avg_opp_GROUND_att	-0.080
R_win_by_TKO_Doctor_Stoppage	0.125	R_avg_opp_LEG_landed	-0.076
B_avg_opp_KD	0.108	R_avg_opp_LEG_att	-0.050
B_avg_opp_TD_pct	0.089	R_win_by_Decision_Majority	-0.044
B_win_by_Submission	0.062	B_StanceOrthodox	-0.031
B_win_by_Decision_Majority	0.055	R_avg_opp_GROUND_landed	-0.020
B_avg_opp_REV	0.038	B_avg_KD	0.000

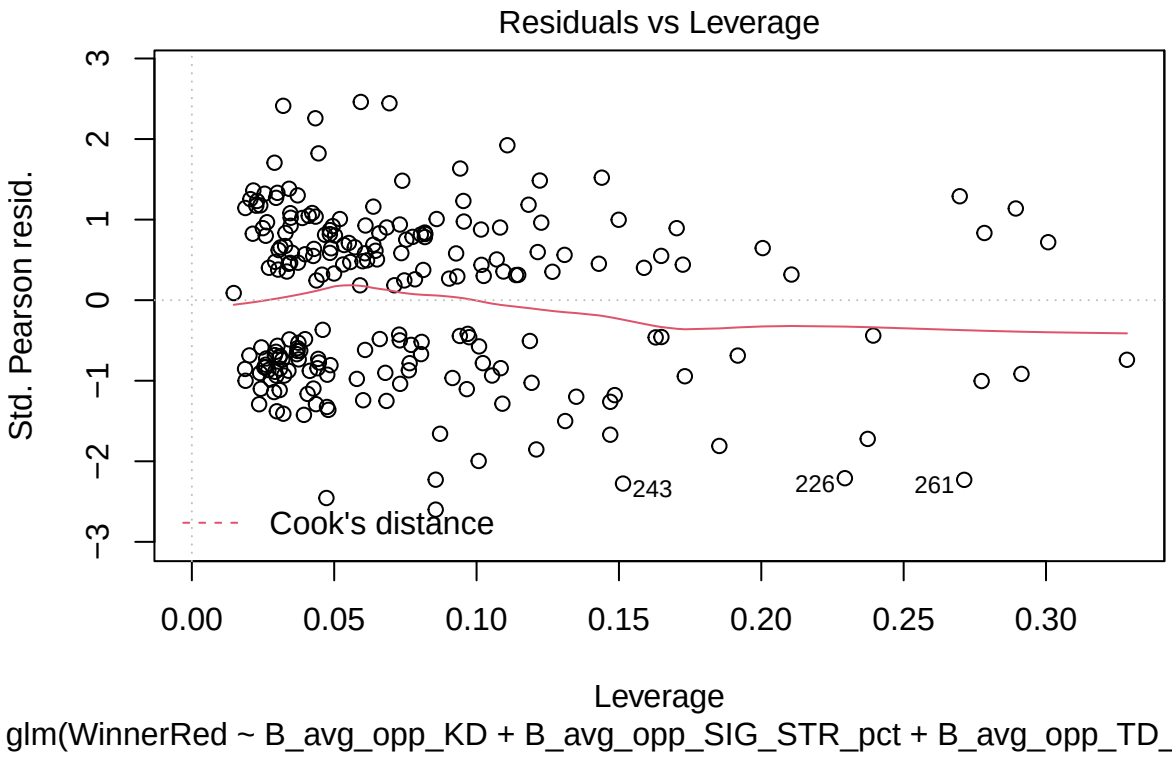


glm(WinnerRed ~ B_avg_opp_KD + B_avg_opp_SIG_STR_pct + B_avg_opp_TD_pct +



glm(WinnerRed ~ B_avg_opp_KD + B_avg_opp_SIG_STR_pct + B_avg_opp_TD_pct +


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
228
121
148
−2
−1
0
1
2
3
4
5
Predicted values
glm(WinnerRed ~ B_avg_opp_KD + B_avg_opp_SIG_STR_pct + B_avg_opp_TD_pct +



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

Full Models

Table 11: Full Class 1 Standardized Coefficients

	s1
(Intercept)	-0.0253
B_avg_KD	0.0534
B_avg_opp_SIG_STR_pct	0.1371
B_avg_opp_REV	0.0050
B_avg_TD_att	-0.0848
B_avg_opp_TD_att	-0.0250
B_avg_opp_HEAD_landed	0.0290
B_avg_opp_LEG_att	-0.0019
B_avg_CLINCH_att	-0.0895
B_losses	-0.0051
B_win_by_Decision_Majority	0.1360
B_win_by_Decision_Split	0.0023
B_win_by_Decision_Unanimous	-0.1862
B_win_by_Submission	-0.0234
R_avg_KD	0.0388
R_avg_opp_SIG_STR_pct	-0.1831
R_avg_TD_pct	-0.0043
R_avg_opp_TD_pct	-0.0624
R_avg_opp_SUB_ATT	-0.0490
R_avg_REV	-0.0568
R_avg_SIG_STR_landed	0.0984
R_avg_BODY_landed	0.1201
R_avg_LEG_att	0.0216
R_avg_opp_LEG_landed	-0.0061
R_avg_opp_CLINCH_att	-0.0743
R_total_title_bouts	0.0018
R_current_win_streak	0.1410
R_losses	-0.0007
R_win_by_KO.TKO	0.0772
R_Reach_cms	0.0049
B_age	0.2156
R_age	-0.2395
B_StanceSwitch	-0.3218

Table 12: Full Class 2 Standardized Coefficients

	s1
(Intercept)	0.0004
B_avg_opp_SIG_STR_pct	0.1140
B_avg_opp_SUB_ATT	-0.0437
B_avg_REV	0.0508
B_avg_TD_att	-0.0596
B_avg_TD_landed	-0.1329
B_avg_CLINCH_landed	0.0159
B_avg_CTRL_time.seconds.	-0.0101
B_total_time_fought.seconds.	-0.0190
B_total_title_bouts	-0.0804
B_current_lose_streak	0.0397
B_longest_win_streak	-0.0226
B_win_by_Decision_Split	0.0107
B_win_by_Submission	-0.0956
B_win_by_TKO_Doctor_Stoppage	0.0458
R_avg_SIG_STR_pct	0.0095
R_avg_SUB_ATT	0.0451
R_avg_opp_REV	0.0144
R_avg_opp_TD_att	0.0060
R_avg_HEAD_landed	0.0344
R_avg_opp_BODY_att	0.0106
R_avg_CLINCH_att	-0.0007
R_avg_opp_CLINCH_att	-0.1079
R_avg_GROUND_att	0.0067
R_avg_opp_CTRL_time.seconds.	-0.1182
R_current_win_streak	0.0645
R_longest_win_streak	0.0307
R_win_by_Decision_Split	-0.1964
R_win_by_Decision_Unanimous	0.0620
R_win_by_Submission	0.0075
B_age	0.3166
R_age	-0.1644
B_StanceOrthodox	0.0365
B_StanceSouthpaw	-0.1994

Table 13: Full Class 3 Standardized Coefficients

	s1
(Intercept)	-0.1783
B_avg_KD	-0.0055
B_avg_opp_KD	0.0147
B_avg_SIG_STR_pct	-0.0112
B_avg_opp_SIG_STR_pct	0.0417
B_avg_TD_pct	0.0016
B_avg_opp_TD_pct	0.0532
B_avg_SUB_ATT	0.0176
B_avg_opp_SUB_ATT	-0.0091
B_avg_REV	-0.0333
B_avg_opp_REV	-0.0308
B_avg_SIG_STR_att	-0.0080
B_avg_SIG_STR_landed	-0.0204
B_avg_opp_SIG_STR_att	0.0106
B_avg_opp_SIG_STR_landed	0.0240
B_avg_TOTAL_STR_att	-0.0009
B_avg_TOTAL_STR_landed	-0.0083
B_avg_opp_TOTAL_STR_att	0.0297
B_avg_opp_TOTAL_STR_landed	0.0480
B_avg_TD_att	-0.0422
B_avg_TD_landed	-0.0319
B_avg_opp_TD_att	-0.0459
B_avg_opp_TD_landed	-0.0083
B_avg_HEAD_att	-0.0055
B_avg_HEAD_landed	-0.0156
B_avg_opp_HEAD_att	0.0192
B_avg_opp_HEAD_landed	0.0340
B_avg_BODY_att	0.0033
B_avg_BODY_landed	-0.0094
B_avg_opp_BODY_att	-0.0006
B_avg_opp_BODY_landed	0.0285
B_avg_LEG_att	-0.0249
B_avg_LEG_landed	-0.0245
B_avg_opp_LEG_att	-0.0385
B_avg_opp_LEG_landed	-0.0324
B_avg_DISTANCE_att	-0.0027
B_avg_DISTANCE_landed	-0.0164
B_avg_opp_DISTANCE_att	-0.0020
B_avg_opp_DISTANCE_landed	0.0011
B_avg_CLINCH_att	-0.0062
B_avg_CLINCH_landed	-0.0038
B_avg_opp_CLINCH_att	0.0276
B_avg_opp_CLINCH_landed	0.0363
B_avg_GROUND_att	-0.0222
B_avg_GROUND_landed	-0.0172
B_avg_opp_GROUND_att	0.0342
B_avg_opp_GROUND_landed	0.0402
B_avg_CTRL_time.seconds.	-0.0336
B_avg_opp_CTRL_time.seconds.	0.0250
B_total_time_fought.seconds.	-0.0324
B_total_rounds_fought	-0.0051
B_total_title_bouts	-0.0187
B_current_win_streak	-0.0369
B_current_lose_streak	0.0224
B_longest_win_streak	-0.0385

Table 14: Full Class 4 Standardized Coefficients

	s1
(Intercept)	0.2134
B_avg_opp_KD	0.1078
B_avg_opp_SIG_STR_pct	0.1385
B_avg_opp_TD_pct	0.0894
B_avg_REV	0.0284
B_avg_opp_REV	0.0384
B_losses	0.0267
B_win_by_Decision_Majority	0.0546
B_win_by_Submission	0.0619
B_Reach_cms	-0.1360
R_avg_opp_LEG_att	-0.0498
R_avg_opp_LEG_landed	-0.0758
R_avg_opp_GROUND_att	-0.0802
R_avg_opp_GROUND_landed	-0.0202
R_win_by_Decision_Majority	-0.0444
R_win_by_TKO_Doctor_Stoppage	0.1255
B_StanceOrthodox	-0.0313