# Final Project

Yifei Yang, Yiming Yuan

## Read in the data

## Introduction and data

Around 300 million individuals worldwide identify themselves as fans of Mixed Martial Arts (MMA), with its popularity peaking in nations such as the United States, the United Kingdom, Brazil, Singapore, and China. The Ultimate Fighting Championship (UFC) is the premier organization in the MMA world. Our motivation is to research on what contributes to the fighters' performance. Thus, the research questions is: Which model we consider effective in predicting fighters' winning ratio? How fighters improve their performance?

Today's data are UFC fighter statistics including:

**wins:** The number of wins the fighter has in their career.

**draws:** The number of draws the fighter has in their career.

**height_cm:** The height of the fighter in centimeters.

**weight_in_kg:** The weight of the fighter in kilograms.

**reach_in_cm:** The reach of the fighter in centimeters.

**stance:** The fighting stance of the fighter (Orthodox/Southpaw/Switch).

**significant_strikes_landed_per_minute:** The average number of significant strikes landed by the fighter per minute.

**significant_striking_accuracy:** The percentage of significant strikes that land successfully for the fighter.

**significant_strikes_absorbed_per_minute:** The average number of significant strikes absorbed by the fighter per minute.

**significant_strike_defence:** The percentage of opponent's significant strikes that the fighter successfully defends.

**average_takedowns_landed_per_15_minutes:** The average number of takedowns landed by the fighter per 15 minutes.

**takedown_accuracy:** The percentage of takedown attempts that are successful for the fighter.

**takedown_defense:** The percentage of opponent's takedown attempts that the fighter successfully defends.

**average_submissions_attempted_per_15_minutes:** The average number of submission attempts made by the fighter per 15 minutes.

Our response variable is wins, predictors are

sources:

https://www.kaggle.com/datasets/aaronfriasr/ufc-fighters-statistics?resource=download

https://www.euronews.com/business/2023/09/27/the-booming-billion-dollar-business-of-combat-sports

*Data Preparation*

```
# A tibble: 6 x 20
  name     nickname  wins losses draws height_cm weight_in_kg reach_in_cm stance
  <chr>    <chr>    <dbl>  <dbl> <dbl>     <dbl>        <dbl>       <dbl> <chr>
1 Robert ~ <NA>         7      0     0      190.         93.0         NA  Ortho~
2 Daniel ~ The Ani~    15     37     0      185.         83.9         NA  <NA>
3 Dan Mol~ <NA>        13      9     0      178.         98.0         NA  <NA>
4 Paul Ru~ <NA>         7      4     0      168.         61.2         NA  <NA>
5 Collin ~ All In       8      2     0      190.         83.9        193. Ortho~
6 Gerald ~ The Fin~     9      7     0      175.         70.3         NA  Ortho~
# i 11 more variables: date_of_birth <date>,
#   significant_strikes_landed_per_minute <dbl>,
#   significant_striking_accuracy <dbl>,
#   significant_strikes_absorbed_per_minute <dbl>,
#   significant_strike_defence <dbl>,
#   average_takedowns_landed_per_15_minutes <dbl>, takedown_accuracy <dbl>,
#   takedown_defense <dbl>, ...
```

## Methodology

1. Predictor selection based on correlation matrix:

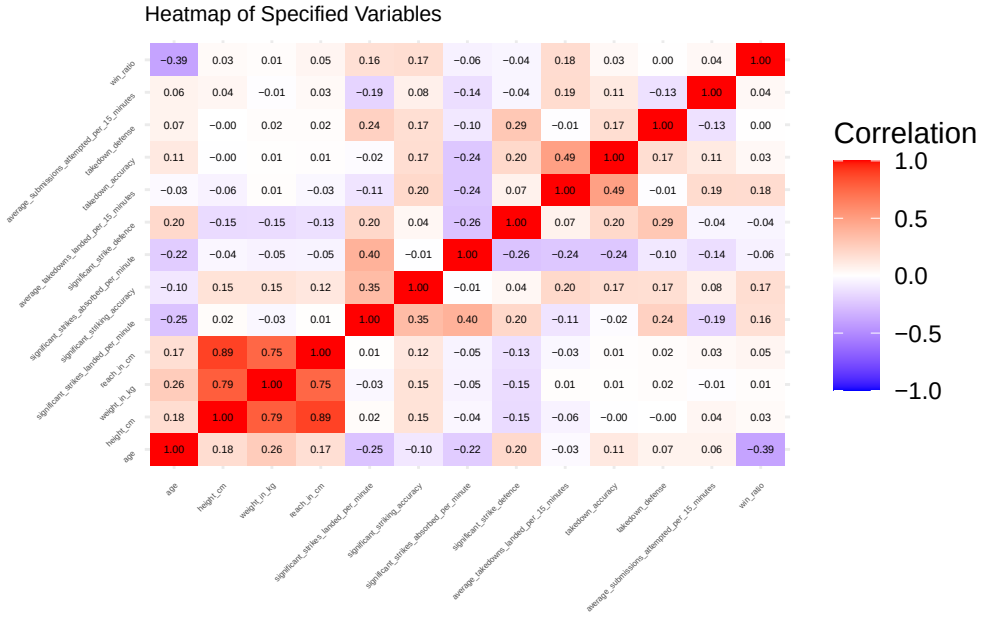
We selected 7 predictors that demonstrated the highest correlation coefficients with the wins ratio for inclusion in the model. They are fighter's age, their stance, the average number of takedowns landed by the fighter per 15 minutes, their striking accuracy, the average number of significant strikes landed by the fighter per minute, the average number of significant strikes absorbed by the fighter per minutem, and the percentage of opponent's significant strikes that the fighter successfully defends.

2. Selection of ordinal model by measurement of prediction accuracy:

We first fit a linear model with with interaction terms and and observed an R-squared value of 0.29, indicating suboptimal predictive performance. Then we change to fitting an ordinal model. We categorize win_ratio into 0-0.33, 0.33-0.67, and 0.67-1, and label them as "Low", "Medium", "High". By creating a confusion matrix, we calculate the test accuracy of 0.71, which concludes that the ordinal model can make more accurate predictions.

## Results

1. *Choose Predictors*




In our analysis, we utilized a heatmap to visually explore and identify the variables most closely associated with fighters' win ratios. The heatmap allowed us to systematically observe the strength and patterns of correlation between various predictors and the win ratio. By presenting the data in this format, we were able to discern which factors are most likely to

influence a fighter's likelihood of winning, thereby providing a foundation for more targeted, in-depth statistical analysis.

Predictors:

age

Stance (category)

average_takendowns_landed_per_15_minutes

Significant_striking_accuracy

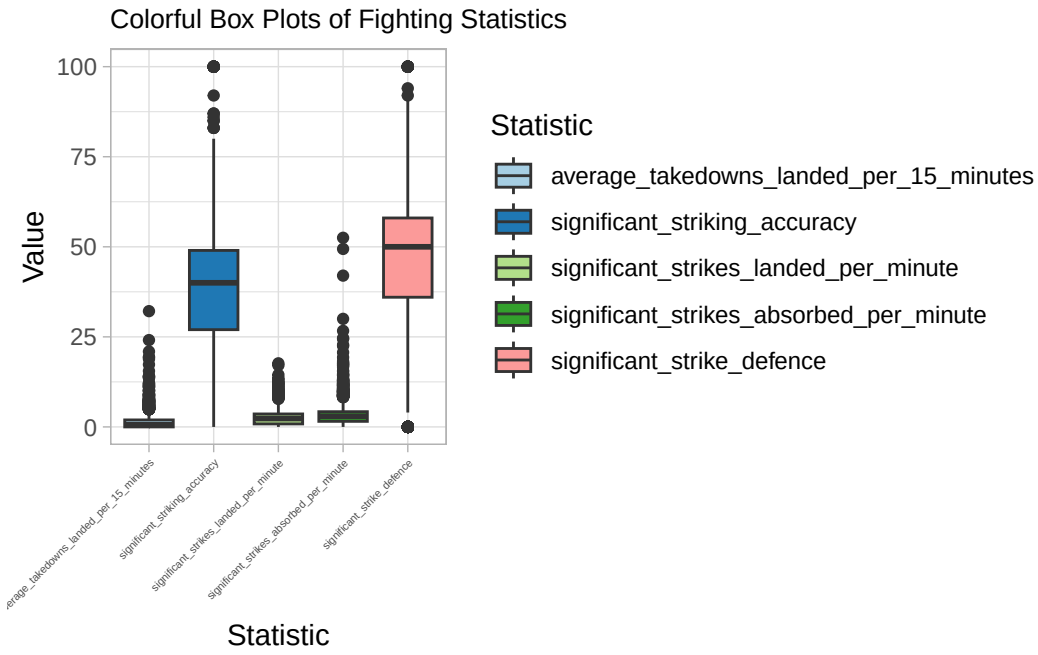
Significant_strikes_landed_per_minute

significant_strikes_absorbed_per_minute

significant_strike_defence

2. *EDA*

We concentrated on the EDA of the response variable to ascertain its distribution characteristics.




3. *Model Fitting*

Our analysis applied an ordinal logistic regression model to explore the relationships between fighters' win ratio groups (categorized as Low, Medium, and High) and a set of predictors.

```
Call:
polr(formula = win_ratio_group ~ age + stance + average_takedowns_landed_per_15_minutes +
    significant_striking_accuracy + significant_strikes_landed_per_minute +
    significant_strikes_absorbed_per_minute + significant_strike_defence,
    data = ufc)

Coefficients:
                                              Value Std. Error  t value
age                                     -0.1151447   0.006670 -17.2641
stanceOrthodox                          -0.3937706   0.816083  -0.4825
stanceSideways                           4.3460381   4.460038   0.9744
stanceSouthpaw                          -0.1067159   0.821314  -0.1299
stanceSwitch                            -0.5320075   0.838556  -0.6344
average_takedowns_landed_per_15_minutes  0.1553674   0.031796   4.8864
significant_striking_accuracy            0.0117169   0.003527   3.3222
significant_strikes_landed_per_minute    0.0675558   0.034174   1.9768
significant_strikes_absorbed_per_minute -0.0528466   0.018069  -2.9248
significant_strike_defence              -0.0007179   0.003301  -0.2175

Intercepts:
            Value    Std. Error t value
Low|Medium   -8.6075   0.9195    -9.3612
Medium|High  -5.0338   0.9050    -5.5622

Residual Deviance: 3562.268
AIC: 3586.268
(1135 observations deleted due to missingness)


        Predicted
Actual    Low Medium High
  Low       2     47    7
  Medium    0    251  620
  High      0    191 1858

[1] "Accuracy:  0.709341397849462"
```

The confusion matrix shows the distribution of actual versus predicted group memberships. The model demonstrates a substantial predictive accuracy of approximately 70.93%, primarily distinguishing well in the higher win ratio groups but showing some limitations in accurately classifying the lower groups.

Thus, it is feasible to find more influential predictor by ordinal model.

Continuous Variable:

- Age shows a negative coefficient (-0.115), indicating that as fighters age, their likelihood of being in a higher win ratio group decreases significantly.
- Average takedowns landed per 15 minutes has a positive coefficient (0.155), suggesting that higher rates of takedowns landed are associated with being in a higher win ratio group.
- Significant striking accuracy has a positive effect (0.012) on the win ratio group, indicating that more accurate strikers tend to be in higher win ratio groups. However, it has a small effect.
- Significant strikes landed per minute also positively affects the win ratio group (coefficient = 0.068), suggesting that fighters who land more strikes per minute tend to have better win ratios.
- Significant strikes absorbed per minute has a negative coefficient (-0.053), indicating that fighters absorbing more strikes per minute are likely to belong to lower win ratio groups.
- Significant strike defence shows a negligible negative coefficient, suggesting minimal impact on win ratio groups.

Categorical variable:

Sideways (coef = 4.3460381): This stance has a significantly positive coefficient, indicating that fighters adopting a Sideways stance have a higher likelihood of being in a higher win ratio group compared to the reference category. However, the statistical significance is marginal (t-value close to 1), thereby it may not be robust across different samples.

Additionally, the coefficients for other stances are all negative, while the t-value indicates that their effect is not statistically significant.

4. *Linear model for comparison*

```
Call:
lm(formula = win_ratio ~ age + stance + average_takedowns_landed_per_15_minutes +
    significant_striking_accuracy + significant_strikes_landed_per_minute +
    significant_strikes_absorbed_per_minute + significant_strike_defence +
    age * average_takedowns_landed_per_15_minutes + age * significant_striking_accuracy +
    age * significant_strikes_absorbed_per_minute + age * significant_strike_defence,
    data = ufc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72177 -0.06487  0.00156  0.07423  0.66218
```

```
Coefficients:
                                             Estimate Std. Error t value
(Intercept)                                 1.130e+00  6.512e-02  17.358
age                                        -1.218e-02  8.557e-04 -14.230
stanceOrthodox                              5.375e-03  4.994e-02   0.108
stanceSideways                              1.853e-01  1.315e-01   1.410
stanceSouthpaw                              9.690e-03  5.017e-02   0.193
stanceSwitch                                1.104e-02  5.080e-02   0.217
average_takedowns_landed_per_15_minutes    -3.342e-03  6.007e-03  -0.556
significant_striking_accuracy              -1.792e-03  7.956e-04  -2.253
significant_strikes_landed_per_minute       7.883e-03  1.656e-03   4.760
significant_strikes_absorbed_per_minute     5.719e-03  4.378e-03   1.306
significant_strike_defence                 -1.763e-03  7.565e-04  -2.331
age:average_takedowns_landed_per_15_minutes 3.698e-04  1.615e-04   2.289
age:significant_striking_accuracy           6.309e-05  1.885e-05   3.347
age:significant_strikes_absorbed_per_minute -3.119e-04  1.141e-04  -2.734
age:significant_strike_defence              4.203e-05  1.773e-05   2.370
                                           Pr(>|t|)
(Intercept)                                 < 2e-16 ***
age                                         < 2e-16 ***
stanceOrthodox                             0.914298
stanceSideways                             0.158761
stanceSouthpaw                             0.846868
stanceSwitch                               0.827888
average_takedowns_landed_per_15_minutes    0.578025
significant_striking_accuracy              0.024348 *
significant_strikes_landed_per_minute      2.03e-06 ***
significant_strikes_absorbed_per_minute    0.191499
significant_strike_defence                 0.019831 *
age:average_takedowns_landed_per_15_minutes 0.022133 *
age:significant_striking_accuracy          0.000826 ***
age:significant_strikes_absorbed_per_minute 0.006295 **
age:significant_strike_defence             0.017840 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1215 on 2961 degrees of freedom
  (1135 observations deleted due to missingness)
Multiple R-squared:  0.2995,    Adjusted R-squared:  0.2962
F-statistic: 90.41 on 14 and 2961 DF,  p-value: < 2.2e-16
```

We also fitted a linear model with with interaction terms and and observed an R-squared value

of 0.29, indicating suboptimal predictive performance.