Final Project

Yifei Yang, Yiming Yuan

Read in the data

```
library(skimr)

library(ggplot2)
library(dplyr)

library(readr)
ufc <- read_csv("data/ufc-fighters-statistics.csv")
```

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 are: How largely do the innate physical attributes affect fighters' performance? What's the most effective stance? How does the way fighters stroke (speed/accuracy/amount/defence success) influence 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

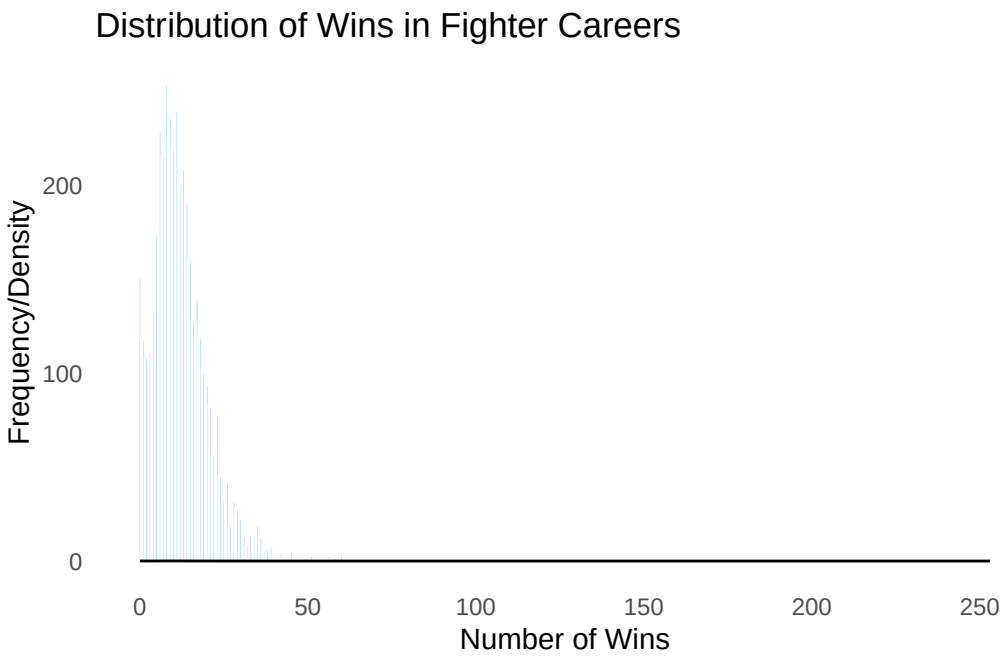
sources:

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

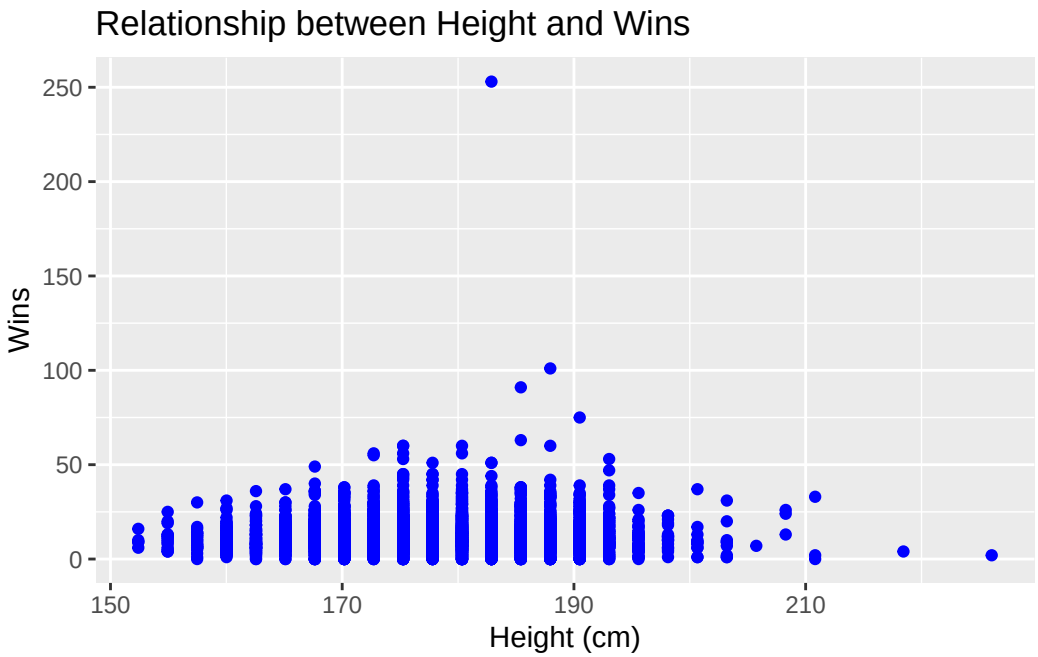
EDA

```
library(ggplot2)

plot <- ggplot(ufc, aes(x = wins)) +
        geom_histogram(binwidth = 1, fill = "skyblue", color = "white") +
        geom_density(alpha = 0.5, fill = "skyblue") +
        labs(x = "Number of Wins", y = "Frequency/Density",
             title = "Distribution of Wins in Fighter Careers") +
        theme_minimal() +
        theme(panel.grid = element_blank())

print(plot)
```



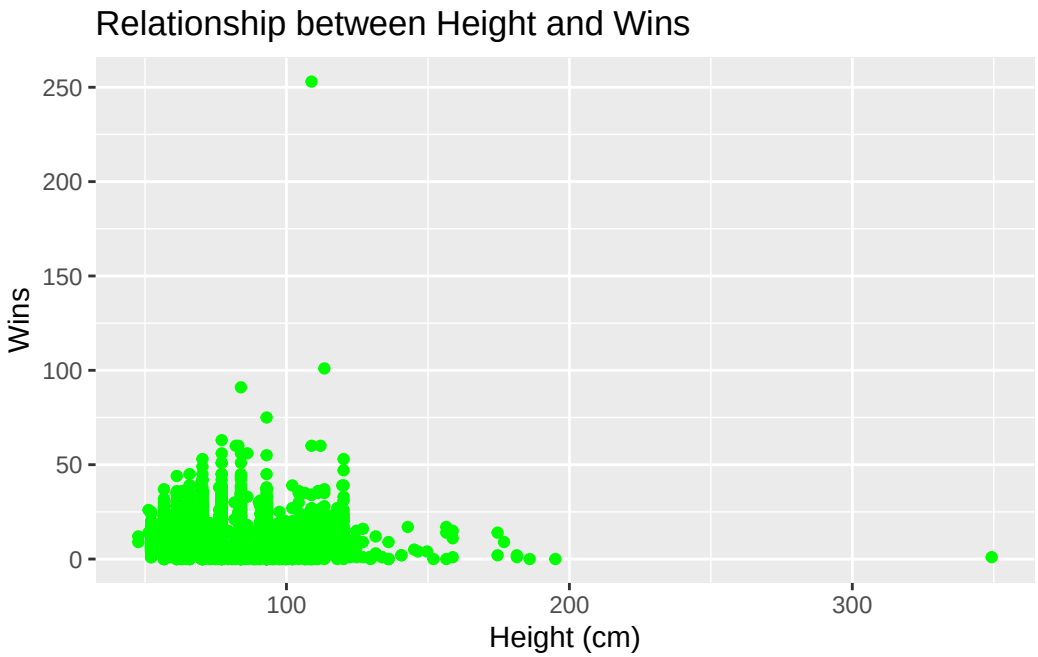

```
plot_height <- ggplot(ufc, aes(x = height_cm, y = wins)) +
               geom_point(color = "blue") +
               labs(x = "Height (cm)", y = "Wins",
                    title = "Relationship between Height and Wins")

print(plot_height)
```

```
Warning: Removed 298 rows containing missing values (`geom_point()`).
```



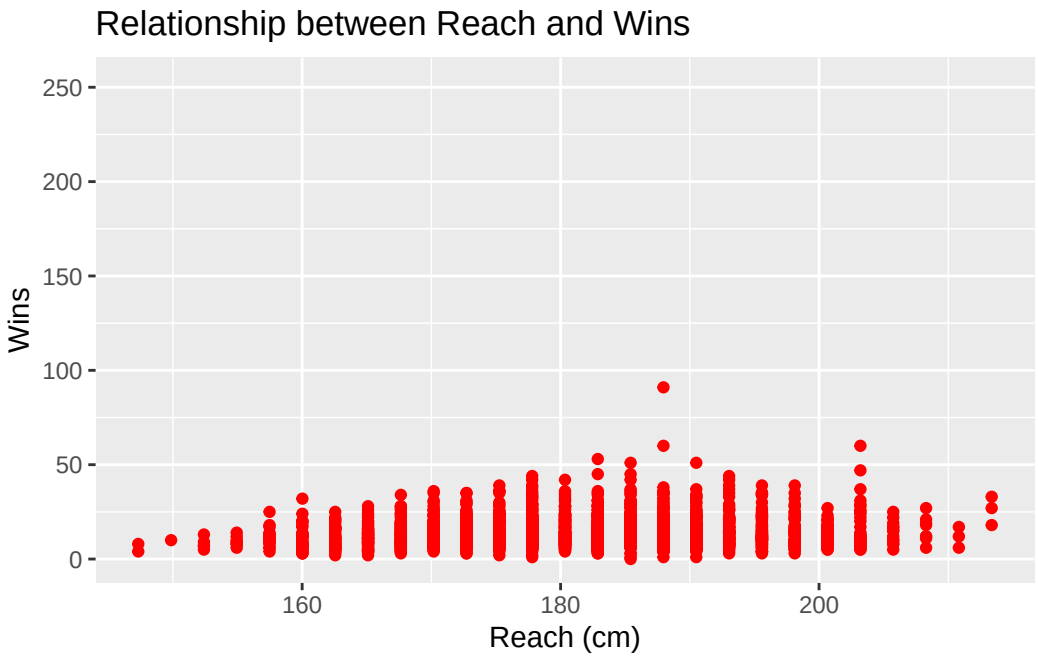

```
plot_weight <- ggplot(ufc, aes(x = weight_in_kg, y = wins)) +
               geom_point(color = "green") +
               labs(x = "Height (cm)", y = "Wins",
                    title = "Relationship between Height and Wins")

print(plot_weight)
```

```
Warning: Removed 87 rows containing missing values (`geom_point()`).
```




```
plot_reach <- ggplot(ufc, aes(x = reach_in_cm, y = wins)) +
              geom_point(color = "red") +
              labs(x = "Reach (cm)", y = "Wins",
                   title = "Relationship between Reach and Wins")
print(plot_reach)
```

```
Warning: Removed 1927 rows containing missing values (`geom_point()`).
```



Data Preparation

```
ufc$win_ratio <- ufc$wins / (ufc$wins + ufc$losses + ufc$draws)

print(head(ufc))
```

```
# A tibble: 6 x 19
  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 10 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>, ...
```

```
skim(ufc)
```

Table 1: Data summary

Name	ufc
Number of rows	4111
Number of columns	19
_____	
Column type frequency:	
character	3
Date	1
numeric	15
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
name	0	1.00	5	27	0	4105	0
nickname	1854	0.55	1	30	0	1784	0
stance	823	0.80	6	11	0	5	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date_of_birth	1135	0.72	1943-01-25	2004-10-08	1986-11-06	2565

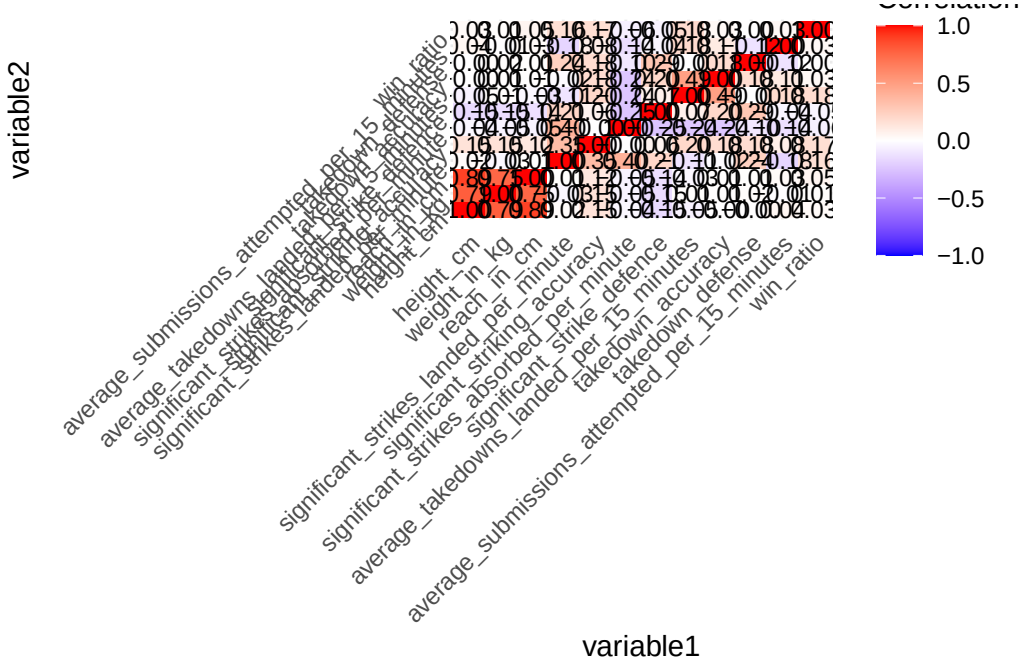
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
wins	0	1.00	12.37	9.37	0.00	7.00	11.00	17.00	253.00	
losses	0	1.00	5.73	5.10	0.00	2.00	5.00	8.00	83.00	
draws	0	1.00	0.26	0.82	0.00	0.00	0.00	0.00	11.00	
height_cm	298	0.93	178.23	8.89	152.40	172.72	177.80	185.42	226.06	
weight_in_kg	87	0.98	77.40	17.98	47.63	65.77	77.11	83.91	349.27	
reach_in_cm	1927	0.53	181.81	10.68	147.32	175.26	182.88	190.50	213.36	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
significant_strikes_landed_per_minute	0	1.00	2.44	1.99	0.00	0.83	2.33	3.60	17.65	
significant_striking_accuracy	0	1.00	35.54	20.40	0.00	27.00	40.00	49.00	100.00	
significant_strikes_absorbed_per_minute	0	1.00	3.15	2.85	0.00	1.55	2.94	4.23	52.50	
significant_strike_defence	0	1.00	42.64	22.32	0.00	36.00	50.00	58.00	100.00	
average_takedowns_landed_per_15_minutes	0	1.00	1.25	1.94	0.00	0.00	0.59	1.94	32.14	
takedown_accuracy	0	1.00	26.30	28.70	0.00	0.00	22.00	45.00	100.00	
takedown_defense	0	1.00	38.96	34.43	0.00	0.00	42.00	66.00	100.00	
average_submissions_attempted_per_15_minutes	0	1.00	0.61	1.51	0.00	0.00	0.00	0.70	21.90	
win_ratio	19	1.00	0.66	0.19	0.00	0.60	0.69	0.78	1.00	

Choose Predictors

```
# Select specified continuous columns
selected_columns <- ufc %>%
  select(height_cm, weight_in_kg, reach_in_cm, significant_strikes_landed_per_minute,
         significant_striking_accuracy, significant_strikes_absorbed_per_minute,
         significant_strike_defence, average_takedowns_landed_per_15_minutes,
         takedown_accuracy, takedown_defense, average_submissions_attempted_per_15_minutes

# Compute the correlation matrix
correlation_matrix <- cor(selected_columns, use = "complete.obs")  # Handles NA by excludi

# Convert the correlation matrix to a long format for ggplot
correlation_data <- as.data.frame(as.table(correlation_matrix))

# Rename columns for clarity
names(correlation_data) <- c("variable1", "variable2", "value")

# Plotting the heatmap
ggplot(correlation_data, aes(x = variable1, y = variable2, fill = value)) +
  geom_tile() +  # This creates the heatmap tiles
  geom_text(aes(label = sprintf("%.2f", value)), color = "black", size = 3) +  # Adds text
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.text.y = element_text(angle = 45, vjust = 1))  # Adjust text alignment if nec
```



Predictors: Stance (category) average_takendowns_landed_per_15 minutes Significant_striking_accuracy Significant_strikes_landed per_minute significant_strikes_absorbed_per_minute significant_strike_defence

Model Fitting

```
ufc$stance <- as.factor(ufc$stance)

model1 <- lm(win_ratio ~ stance + average_takedowns_landed_per_15_minutes + significant_st
             significant_strikes_landed_per_minute + significant_strikes_absorbed_per_minut
             data = ufc)


summary(model1)
```

```
Call:
lm(formula = win_ratio ~ 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)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85251 -0.06706  0.00832  0.08643  0.50832

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              0.4481633  0.0590408   7.591 4.12e-14
stanceOrthodox                           0.0435211  0.0587530   0.741    0.459
stanceSideways                          -0.0893852  0.1070817  -0.835    0.404
stanceSouthpaw                           0.0484787  0.0590289   0.821    0.412
stanceSwitch                             0.0895293  0.0597436   1.499    0.134
average_takedowns_landed_per_15_minutes  0.0113384  0.0015128   7.495 8.48e-14
significant_striking_accuracy            0.0017843  0.0002011   8.875  < 2e-16
significant_strikes_landed_per_minute    0.0199915  0.0018413  10.857  < 2e-16
significant_strikes_absorbed_per_minute -0.0016777  0.0010519  -1.595    0.111
significant_strike_defence               0.0009887  0.0001818   5.439 5.76e-08

(Intercept)                             ***
stanceOrthodox
stanceSideways
stanceSouthpaw
stanceSwitch
average_takedowns_landed_per_15_minutes ***
significant_striking_accuracy           ***
significant_strikes_landed_per_minute   ***
significant_strikes_absorbed_per_minute
significant_strike_defence              ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.155 on 3278 degrees of freedom
  (823 observations deleted due to missingness)
Multiple R-squared:  0.1899,    Adjusted R-squared:  0.1877
F-statistic: 85.38 on 9 and 3278 DF,  p-value: < 2.2e-16
```

Categorical variable:

0.0435211 represents the difference in average win ratio for fighters with Orthodox stance, compared with fighters with Open stance, while holding the other predictors constant.

-0.0893852 represents the difference in average win ratio for fighters with Sideways stance, compared with fighters with Open stance, while holding the other predictors constant.

0.0484787 represents the difference in average win ratio for fighters with Southpaw stance, compared with fighters with Open stance, while holding the other predictors constant.

0.0895293 represents the difference in average win ratio for fighters with Switch stance, compared with fighters with Open stance, while holding the other predictors constant.

We notice that the p-value for Stance variables are all less than 0.05, and thus, there is insufficient evidence to suggest a linear relationship between fighter's stance and wins ratio at 0.05 significance level, while controlling for other predictors

Continuous variable:

While holding the other predictors constant, when the average number of takedowns landed by the fighter per 15 minutes increases by 1 time, the average win ratio of the fighter will increase by 0.0113384.

While holding the other predictors constant, when the significant striking accuracy increases by 1%, the average win ratio of the fighter will increase by 0.0017843.

While holding the other predictors constant, when the average number of significant strikes landed by the fighter per minute increases by 1 time, the average win ratio of the fighter will increase by 0.0199915.

While holding the other predictors constant, when the average number of significant strikes absorbed by the fighter per minute increases by 1 time, the average win ratio of the fighter will decrease by 0.0016777.

While holding the other predictors constant, when the opponent's significant strikes that the fighter successfully defends increases by 1%, the average win ratio of the fighter will increase by 0.0009887.

```r
ufc$win_ratio_binary <- ifelse(ufc$win_ratio > 0.5, 1, 0)

model2 <- glm(win_ratio_binary ~ stance + average_takedowns_landed_per_15_minutes +
                        significant_striking_accuracy + significant_strikes_landed_per_minut
                        significant_strikes_absorbed_per_minute + significant_strike_defence
                        family = binomial, data = ufc)

summary(model2)
```

```
Call:
glm(formula = win_ratio_binary ~ stance + average_takedowns_landed_per_15_minutes +
    significant_striking_accuracy + significant_strikes_landed_per_minute +
    significant_strikes_absorbed_per_minute + significant_strike_defence,
    family = binomial, data = ufc)
```

```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -0.559828   0.938009  -0.597 0.550624
stanceOrthodox                          0.423837   0.936999   0.452 0.651028
stanceSideways                         -0.542182   1.573415  -0.345 0.730403
stanceSouthpaw                          0.354448   0.945138   0.375 0.707644
stanceSwitch                            1.366299   1.002463   1.363 0.172901
average_takedowns_landed_per_15_minutes 0.155218   0.047682   3.255 0.001133
significant_striking_accuracy           0.016132   0.004273   3.775 0.000160
significant_strikes_landed_per_minute   0.438657   0.059543   7.367 1.74e-13
significant_strikes_absorbed_per_minute -0.010539   0.018663  -0.565 0.572285
significant_strike_defence              0.013010   0.003536   3.680 0.000234

(Intercept)
stanceOrthodox
stanceSideways
stanceSouthpaw
stanceSwitch
average_takedowns_landed_per_15_minutes **
significant_striking_accuracy           ***
significant_strikes_landed_per_minute   ***
significant_strikes_absorbed_per_minute
significant_strike_defence              ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2288.4  on 3287  degrees of freedom
Residual deviance: 1859.4  on 3278  degrees of freedom
  (823 observations deleted due to missingness)
AIC: 1879.4

Number of Fisher Scoring iterations: 6
```

We impose a threshold of 0.5 for wins ration: when it's larger than 0.5, we consider it Satisfactory Performance, and when it's less than 0.5, we consider it Unsatisfactory Performance.