



BLUE vs RED

승부예측



박지은 오민석 이정민 장은조

START

INDEX



1

UFC / 초기 Data 소개

2

EDA / 전처리 및 결측치 처리

3

Modeling 및 결과

4

한계점



INDEX 1
UFC / 초기 Data 소개



INDEX 2
EDA / 전처리 및 결측치 처리



INDEX 3
Modeling 및 결과



INDEX 4
한계점



UFC / 초기 Data 소개

UFC



UFC

KO

바닥 두들기기, 합법적 타격으로 쓰러져 10초 이상 일어나지 못할 시

TKO

선수가 부상을 당하거나 전의를 상실해 심판이 경기를 중단할 시

판정

항복 받는 서브미션으로 경기가 끝나지 않고 모든 라운드를 마치게 되면 3명의 주심에 의해 판정을 내려 승자를 결정

RAW DATA



[데이터 소개]

R_fighter	B_fighter	Referee	date	location	Winner	title_bout	weight_class	no_of_rounds	B_current_lose_streak	B_current_win_streak	B_draw
Henry Cejudo	Marlon Moraes	Marc Goddard	2019-06-08	Chicago, Illinois, USA	Red	TRUE	Bantamweight	5	0	4	0
Valentina Shevchenko	Jessica Eye	Robert Madrigal	2019-06-08	Chicago, Illinois, USA	Red	TRUE	Women's Flyweight	5	0	3	0
Tony Ferguson	Donald Cerrone	Dan Miragliotta	2019-06-08	Chicago, Illinois, USA	Red	FALSE	Lightweight	3	0	3	0
Jimmie Rivera	Petr Yan	Kevin MacDonald	2019-06-08	Chicago, Illinois, USA	Blue	FALSE	Bantamweight	3	0	4	0
Tai Tuivasa	Blagoy Ivanov	Dan Miragliotta	2019-06-08	Chicago, Illinois, USA	Blue	FALSE	Heavyweight	3	0	1	0
Tatiana Suarez	Nina Ansaroff	Robert Madrigal	2019-06-08	Chicago, Illinois, USA	Red	FALSE	Women's Strawweight	3	0	4	0
Aljamain Sterling	Pedro Munhoz	Marc Goddard	2019-06-08	Chicago, Illinois, USA	Red	FALSE	Bantamweight	3	0	3	0
Karolina Kowalkiewicz	Alexa Grasso	Kevin MacDonald	2019-06-08	Chicago, Illinois, USA	Blue	FALSE	Women's Strawweight	3	1	0	0
Ricardo Lamas	Calvin Kattar	Dan Miragliotta	2019-06-08	Chicago, Illinois, USA	Blue	FALSE	Featherweight	3	0	1	0
Yan Xiaonan	Angela Hill	Robert Madrigal	2019-06-08	Chicago, Illinois, USA	Red	FALSE	Women's Strawweight	3	0	1	0
Bevon Lewis	Darren Stewart	Marc Goddard	2019-06-08	Chicago, Illinois, USA	Blue	FALSE	Middleweight	3	1	0	0
Eddie Wineland	Grigori Popov	Kevin MacDonald	2019-06-08	Chicago, Illinois, USA	Red	FALSE	Bantamweight	3	0	0	0
Katlyn Chookagian	Joanne Calderwood	Dan Miragliotta	2019-06-08	Chicago, Illinois, USA	Red	FALSE	Women's Flyweight	3	0	2	0
Alexander Gustafsson	Anthony Smith	Marc Goddard	2019-06-01	Stockholm, Sweden	Blue	FALSE	Light Heavyweight	5	1	0	0
Jimi Manuwa	Aleksandar Rakic	Kevin Sataki	2019-06-01	Stockholm, Sweden	Blue	FALSE	Light Heavyweight	3	0	3	0
Makwan Amirkhani	Chris Fishgold	Marc Goddard	2019-06-01	Stockholm, Sweden	Red	FALSE	Featherweight	3	0	1	0
Damir Hadzovic	Christos Giagos	Rich Mitchell	2019-06-01	Stockholm, Sweden	Blue	FALSE	Lightweight	3	0	1	0
Daniel Teymur	Sung Bin Jo	Rebin Saber	2019-06-01	Stockholm, Sweden	Red	FALSE	Featherweight	3	0	0	0
Rostem Akman	Sergey Khandozhko	Leon Roberts	2019-06-01	Stockholm, Sweden	Blue	FALSE	Welterweight	3	0	0	0
Tonya Evinger	Lina Lansberg	Kevin Sataki	2019-06-01	Stockholm, Sweden	Blue	FALSE	Women's Bantamweight	3	1	0	0
Stevie Ray	Leonardo Santos	Marc Goddard	2019-06-01	Stockholm, Sweden	Blue	FALSE	Lightweight	3	0	4	0
Nick Hein	Frank Camacho	Leon Roberts	2019-06-01	Stockholm, Sweden	Blue	FALSE	Lightweight	3	2	0	0
Bea Malecki	Duda Santana		2019-06-01	Stockholm, Sweden	Red	FALSE	Women's Bantamweight	3	0	0	0
Darko Stosic	Devin Clark	Kevin Sataki	2019-06-01	Stockholm, Sweden	Blue	FALSE	Light Heavyweight	3	1	0	0
Joel Alvarez	Danilo Belluardo	Rebin Saber	2019-06-01	Stockholm, Sweden	Red	FALSE	Lightweight	3	0	0	0
Rafael Dos Anjos	Kevin Lee	Dan Miragliotta	2019-05-18	Rochester, New York, USA	Red	FALSE	Welterweight	5	1	0	0
Antonio Carlos Junior	Ian Heinisch	odd Ronald Anders	2019-05-18	Rochester, New York, USA	Blue	FALSE	Middleweight	3	0	1	0
Megan Anderson	Felicia Spencer	Keith Peterson	2019-05-18	Rochester, New York, USA	Blue	FALSE	Women's Featherweight	3	0	0	0
Vicente Luque	Derrick Krantz	Dan Miragliotta	2019-05-18	Rochester, New York, USA	Red	FALSE	Welterweight	3	0	0	0

변수 145개

데이터 5144개



INDEX 1
UFC / 초기 Data 소개



INDEX 2
EDA / 전처리 및 결측치 처리



INDEX 3
Modeling 및 결과



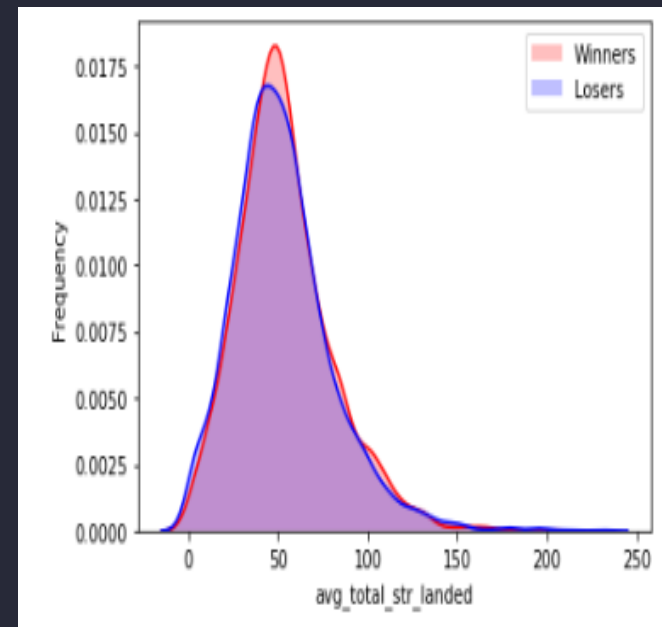
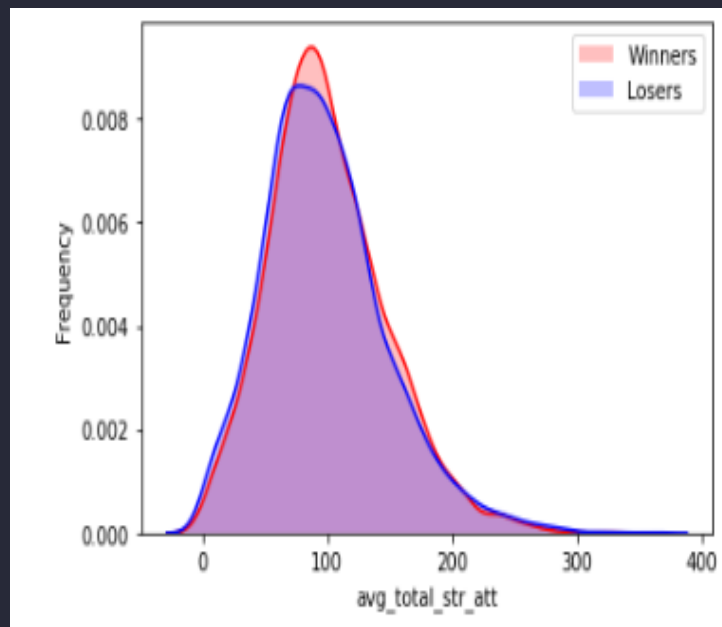
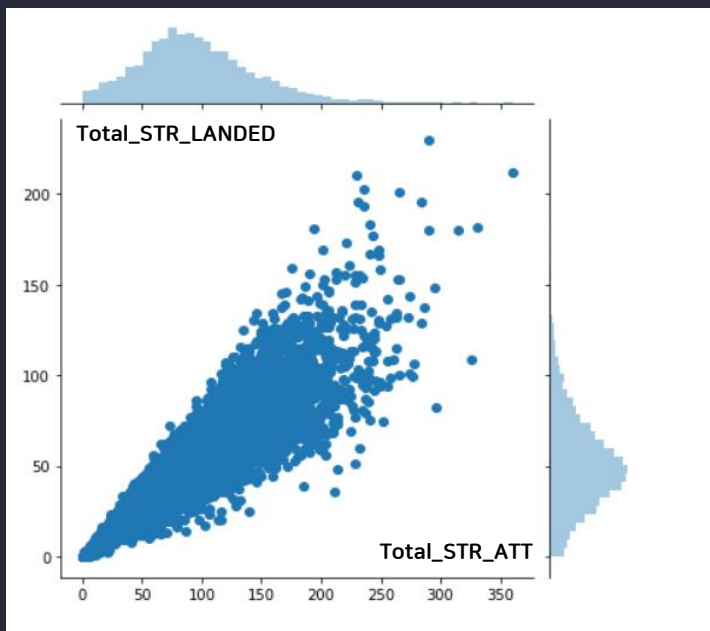
INDEX 4
한계점

UFC



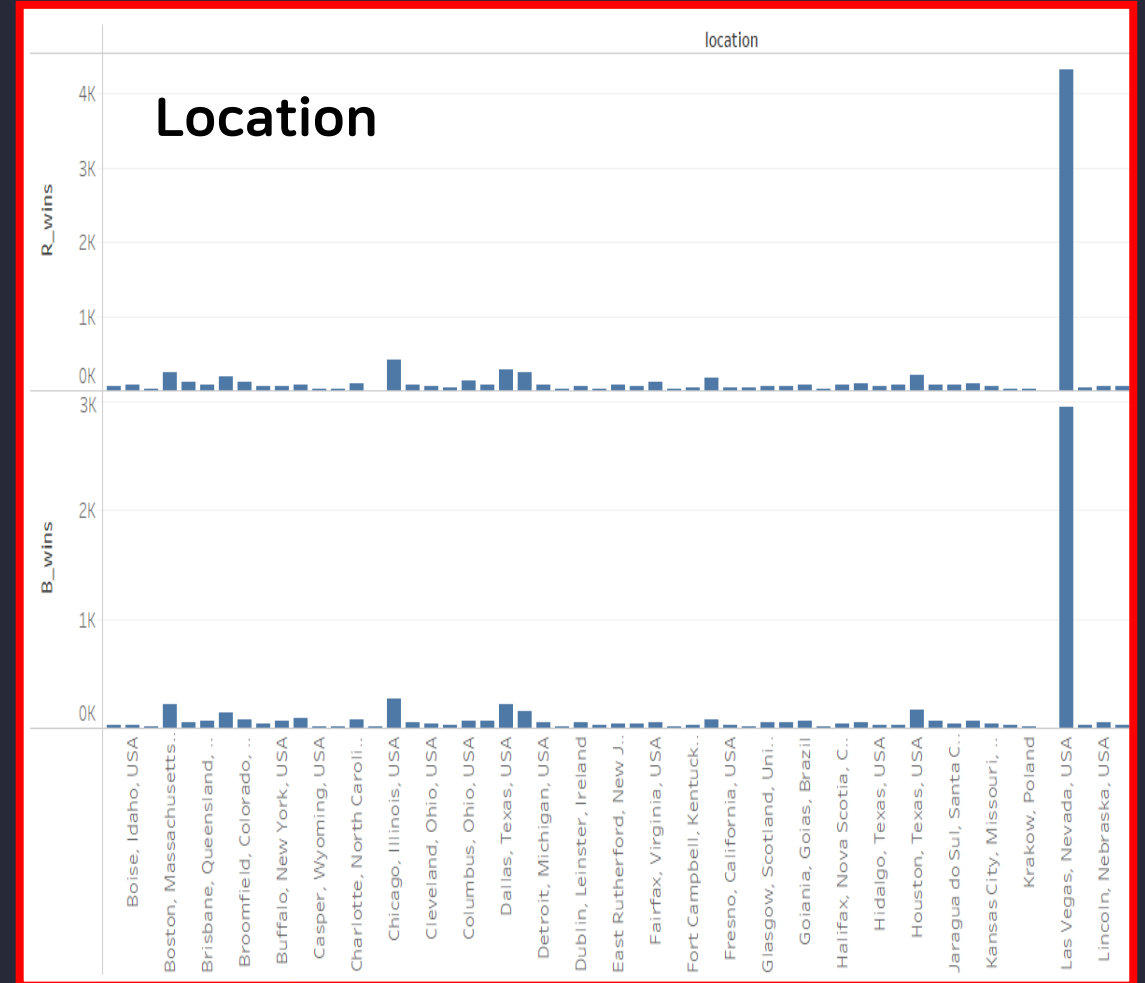
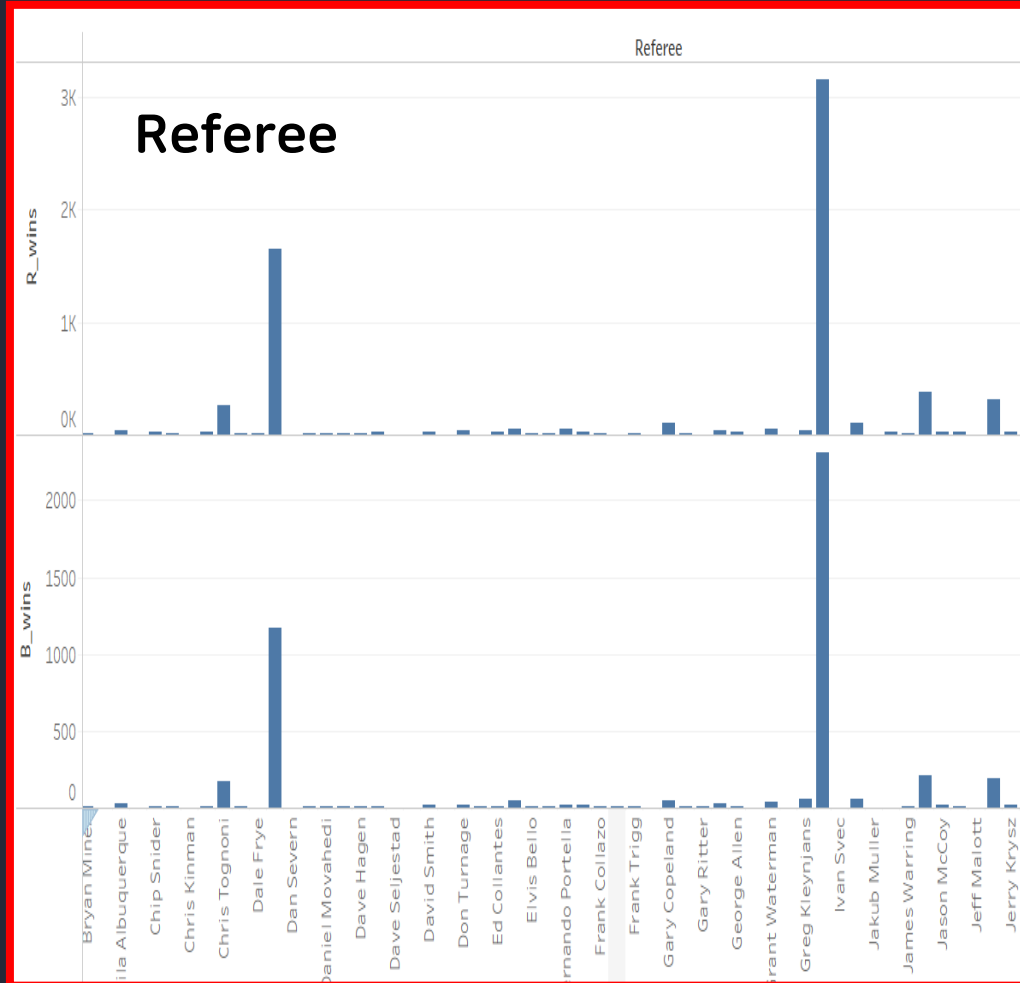
EDA / 전처리 및 결측치 처리

EDA



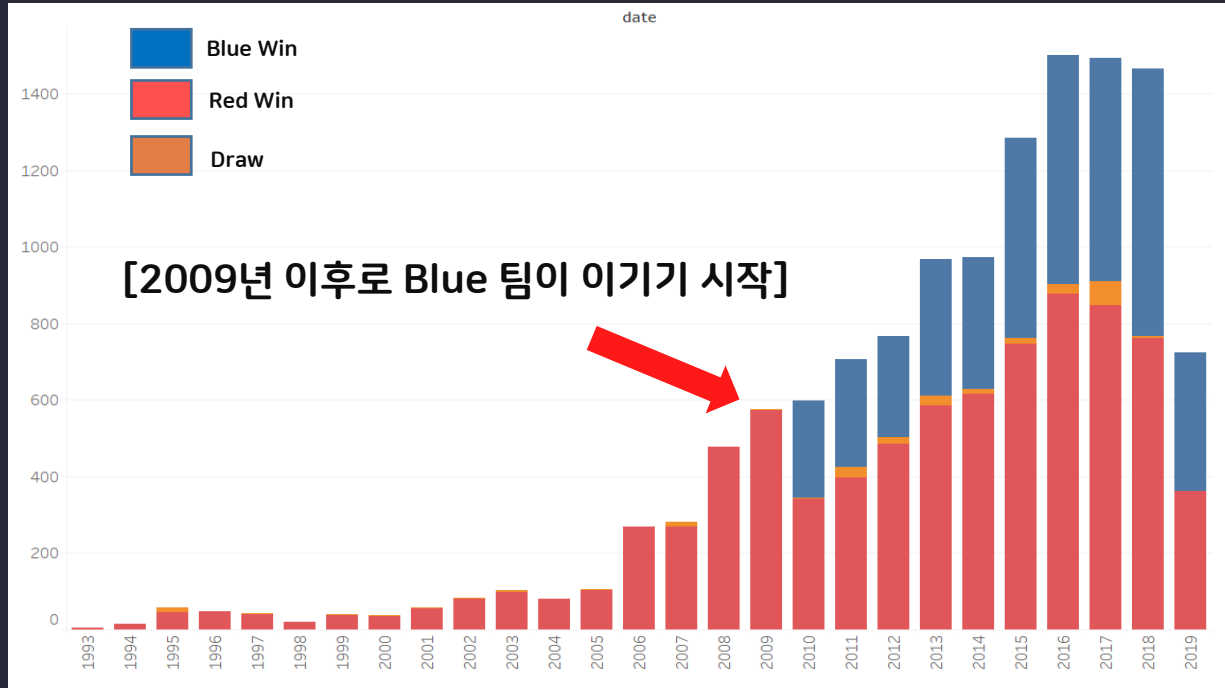
승부를 결정 짓는 유의미한 변수를 찾아내지 못함

EDA



Referee, Location 승부에 큰 영향을 미치지 않는다고 판단

EDA



Committee Report on Unified Rules for MMA

SUMMARY REPORT

Discussion and Review of UNIFIED RULES OF MIXED MARTIAL ARTS

On July 30, 2009, A motion was made by Bill Colbert, (UT) at the annual meeting of the Association of Boxing Commissions to adopt these rule as the Unified Rules of Mixed Martial Arts with the understanding that any commission can require them more stringent. Motion was seconded by Joel Campuzano, (IL) Motion Pass Unanimously.

[2009년 7월 30일 공식적인 규칙이 생김]

결측치 특징(1)



변수	결측치개수	변수	결측치개수
R_fighter	0	B_win_by_KO.TKO	0
B_fighter	0	B_win_by_Submission	0
Referee	0	B_win_by_TKO_Doctor_Stoppage	0
date	0	B_wins	0
location	0	B_Stance	0
Winner	0	B_Weight_lbs	0
title_bout	0	R_current_lose_streak	0
weight_class	0	R_current_win_streak	0
no_of_rounds	0	R_draw	0
B_current_lose_streak	0	R_longest_win_streak	0
B_current_win_streak	0	R_losses	0
B_draw	0	R_total_rounds_fought	0
B_longest_win_streak	0	R_total_title_bouts	0
B_losses	0	R_win_by_Decision_Majority	0
B_total_rounds_fought	0	R_win_by_Decision_Split	0
B_total_title_bouts	0	R_win_by_Decision_Unanimous	0
B_win_by_Decision_Majority	0	R_win_by_KO.TKO	0
B_win_by_Decision_Split	0	R_win_by_Submission	0
B_win_by_Decision_Unanimous	0	R_win_by_TKO_Doctor_Stoppage	0
R_Stance	0	R_wins	0

[42개의 변수 : 결측치 없음]

변수	결측치개수
R_Weight_lbs	1
B_Height_cms	2
R_Height_cms	2
R_age	2
B_age	9
R_Reach_cms	41
B_Reach_cms	110

[7개의 변수 : 결측치 있음]

- 단순 누락으로 인한 결측치로 판단
- Weight_lbs , Height, age : by searching
- Reach : Height와 Reach의 상관관계 89%
: by linear regression

결측치 특징(2)



변수	결측치개수	변수	결측치개수
R_avg_BODY_att	380	R_avg_opp_DISTANCE_att	380
R_avg_BODY_landed	380	R_avg_opp_DISTANCE_landed	380
R_avg_CLINCH_att	380	R_avg_opp_GROUND_att	380
R_avg_CLINCH_landed	380	R_avg_opp_GROUND_landed	380
R_avg_DISTANCE_att	380	R_avg_opp_HEAD_att	380
R_avg_DISTANCE_landed	380	R_avg_opp_HEAD_landed	380
R_avg_GROUND_att	380	R_avg_opp_KD	380
R_avg_GROUND_landed	380	R_avg_opp_LEG_att	380
R_avg_HEAD_att	380	R_avg_opp_LEG_landed	380
R_avg_HEAD_landed	380	R_avg_opp_PASS	380
R_avg_KD	380	R_avg_opp_REV	380
R_avg_LEG_att	380	R_avg_opp_SIG_STR_att	380
R_avg_LEG_landed	380	R_avg_opp_SIG_STR_landed	380
R_avg_PASS	380	R_avg_opp_SIG_STR_pct	380
R_avg_REV	380	R_avg_opp_SUB_ATT	380
R_avg_SIG_STR_att	380	R_avg_opp_TD_att	380
R_avg_SIG_STR_landed	380	R_avg_opp_TD_landed	380
R_avg_SIG_STR_pct	380	R_avg_opp_TD_pct	380
R_avg_SUB_ATT	380	R_avg_opp_TOTAL_STR_att	380
R_avg_TD_att	380	R_avg_opp_TOTAL_STR_landed	380
R_avg_TD_landed	380	R_total_time_fought.seconds.	380
R_avg_TD_pct	380	R_avg_opp_BODY_att	380
R_avg_TOTAL_STR_att	380	R_avg_opp_BODY_landed	380
R_avg_TOTAL_STR_landed	380	R_avg_opp_CLINCH_att	380
		R_avg_opp_CLINCH_landed	380

[49개의 변수 : 각 변수별로 380개의 결측치 존재]

변수	결측치개수	변수	결측치개수
B_avg_BODY_att	877	B_avg_opp_BODY_landed	877
B_avg_BODY_landed	877	B_avg_opp_CLINCH_att	877
B_avg_CLINCH_att	877	B_avg_opp_CLINCH_landed	877
B_avg_CLINCH_landed	877	B_avg_opp_DISTANCE_att	877
B_avg_DISTANCE_att	877	B_avg_opp_DISTANCE_landed	877
B_avg_DISTANCE_landed	877	B_avg_opp_GROUND_att	877
B_avg_GROUND_att	877	B_avg_opp_GROUND_landed	877
B_avg_GROUND_landed	877	B_avg_opp_HEAD_att	877
B_avg_HEAD_att	877	B_avg_opp_HEAD_landed	877
B_avg_HEAD_landed	877	B_avg_opp_KD	877
B_avg_KD	877	B_avg_opp_LEG_att	877
B_avg_LEG_att	877	B_avg_opp_LEG_landed	877
B_avg_LEG_landed	877	B_avg_opp_PASS	877
B_avg_PASS	877	B_avg_opp_REV	877
B_avg_REV	877	B_avg_opp_SIG_STR_att	877
B_avg_SIG_STR_att	877	B_avg_opp_SIG_STR_landed	877
B_avg_SIG_STR_landed	877	B_avg_opp_SIG_STR_pct	877
B_avg_SIG_STR_pct	877	B_avg_opp_SUB_ATT	877
B_avg_SUB_ATT	877	B_avg_opp_TD_att	877
B_avg_TD_att	877	B_avg_opp_TD_landed	877
B_avg_TD_landed	877	B_avg_opp_TD_pct	877
B_avg_TD_pct	877	B_avg_opp_TOTAL_STR_att	877
B_avg_TOTAL_STR_att	877	B_avg_opp_TOTAL_STR_landed	877
B_avg_TOTAL_STR_landed	877	B_total_time_fought.seconds.	877
B_avg_opp_BODY_att	877		

[49개의 변수 : 각 변수별로 877개의 결측치 존재]

결측치 특징(2)



```
> nrow(df[df$B_total_rounds_fought==0,])  
[1] 877  
> nrow(df[df$R_total_rounds_fought==0,])  
[1] 380
```

> total_rounds_fought : 선수가 뒀 경기의 누적된 라운드

> total_rounds_fought ==0 : 처음 출전한 선수

“처음 출전한 선수들은 opp, avg 같은 49개의 누적변수들이 결측치”

각 선수의 첫경기 데이터를 제외한 데이터들의 평균을 이용해 결측치 처리

> 모든 데이터들을 가지고 첫 경기 데이터를 예측하기에는 변동을 고려하기 어려움

> 그래서 첫 경기와 가장 가까운 시기에 한 두경기의 데이터의 평균을 이용해 결측치 처리

결측치 처리



변수	결측치개수	변수	결측치개수	변수	결측치개수	변수	결측치개수	변수	결측치개수
R_fighter	0	B_avg_TD_att	0	B_avg_opp_TOTAL_STR_landed	0	R_avg_LEG_att	0	R_avg_TD_att	0
B_fighter	0	B_avg_TD_landed	0	B_total_rounds_fought	0	R_avg_LEG_landed	0	R_avg_TD_landed	0
Winner	0	B_avg_TD_pct	0	B_total_time_fought.seconds	0	R_avg_PASS	0	R_avg_TD_pct	0
date	0	B_avg_TOTAL_STR_att	0	B_total_title_bouts	0	R_avg_REV	0	R_avg_TOTAL_STR_att	0
title_bout	0	B_avg_TOTAL_STR_landed	0	B_win_by_Decision_Majority	0	R_avg_SIG_STR_att	0	R_avg_TOTAL_STR_landed	0
weight_class	0	B_longest_win_streak	0	B_win_by_Decision_Split	0	R_avg_SIG_STR_landed	0	R_longest_win_streak	0
no_of_rounds	0	B_losses	0	B_win_by_Decision_Unanimous	0	R_avg_SIG_STR_pct	0	R_losses	0
B_current_lose_streak	0	B_avg_opp_BODY_att	0	B_win_by_KO.TKO	0	R_avg_SUB_ATT	0	R_avg_opp_BODY_att	0
B_current_win_streak	0	B_avg_opp_BODY_landed	0	B_win_by_Submission	0	R_avg_TD_att	0	R_avg_opp_BODY_landed	0
B_avg_BODY_att	0	B_avg_opp_CLINCH_att	0	B_win_by_TKO_Doctor_Stoppage	0	R_avg_TD_landed	0	R_total_rounds_fought	0
B_avg_BODY_landed	0	B_avg_opp_CLINCH_landed	0	B_wins	0	R_avg_TD_pct	0	R_total_time_fought.seconds	0
B_avg_CLINCH_att	0	B_avg_opp_DISTANCE_att	0	B_Stance	0	R_avg_TOTAL_STR_att	0	R_total_title_bouts	0
B_avg_CLINCH_landed	0	B_avg_opp_DISTANCE_landed	0	B_Height_cms	0	R_avg_TOTAL_STR_landed	0	R_win_by_Decision_Majority	0
B_avg_DISTANCE_att	0	B_avg_opp_GROUND_att	0	B_Reach_cms	0	R_longest_win_streak	0	R_win_by_Decision_Split	0
B_avg_DISTANCE_landed	0	B_avg_opp_GROUND_landed	0	B_Weight_lbs	0	R_losses	0	R_win_by_Decision_Unanimous	0
B_avg_GROUND_att	0	B_avg_opp_HEAD_att	0	B_age	0	R_avg_opp_BODY_att	0	R_win_by_KO.TKO	0
B_avg_GROUND_landed	0	B_avg_opp_HEAD_landed	0	B_total	0	R_avg_opp_BODY_landed	0	R_win_by_Submission	0
B_avg_HEAD_att	0	B_avg_opp_KD	0	R_current_lose_streak	0	R_avg_opp_CLINCH_att	0	R_win_by_TKO_Doctor_Stoppage	0
B_avg_HEAD_landed	0	B_avg_opp_LEG_att	0	R_current_win_streak	0	R_avg_opp_CLINCH_landed	0	R_wins	0
B_avg_KD	0	B_avg_opp_LEG_landed	0	R_avg_BODY_att	0	R_avg_opp_DISTANCE_att	0	R_Stance	0
B_avg_LEG_att	0	B_avg_opp_PASS	0	R_avg_BODY_landed	0	R_avg_opp_DISTANCE_landed	0	R_Height_cms	0
B_avg_LEG_landed	0	B_avg_opp_REV	0	R_avg_CLINCH_att	0	R_avg_opp_GROUND_att	0	R_Reach_cms	0
B_avg_PASS	0	B_avg_opp_SIG_STR_att	0	R_avg_CLINCH_landed	0	R_avg_LEG_att	0	R_Weight_lbs	0
B_avg_REV	0	B_avg_opp_SIG_STR_landed	0	R_avg_DISTANCE_att	0	R_avg_LEG_landed	0	R_age	0
B_avg_SIG_STR_att	0	B_avg_opp_SIG_STR_pct	0	R_avg_DISTANCE_landed	0	R_avg_PASS	0	R_total	0
B_avg_SIG_STR_landed	0	B_avg_opp_SUB_ATT	0	R_avg_GROUND_att	0	R_avg_REV	0		
B_avg_SIG_STR_pct	0	B_avg_opp_TD_att	0	R_avg_GROUND_landed	0	R_avg_SIG_STR_att	0		
B_avg_SUB_ATT	0	B_avg_opp_TD_landed	0	R_avg_HEAD_att	0	R_avg_SIG_STR_landed	0		
		B_avg_opp_TD_pct	0	R_avg_HEAD_landed	0	R_avg_SIG_STR_pct	0		
		B_avg_opp_TOTAL_STR_att	0	R_avg_KD	0	R_avg_SUB_ATT	0		



INDEX 1
UFC / 초기 Data 소개



INDEX 2
EDA / 전처리 및 결측치 처리



INDEX 3
Modeling 및 결과



INDEX 4
한계점

UFC



Modeling 및 결과

Modeling 및 결과_ 모델링 이전에 준비



1. R_fighter, B_fighter, date 변수 삭제 & 무승부 삭제
2. Winner, title_bout 변수에 대해서 라벨 인코딩
3. weight_class, B_Stance, R_Stance에 대해 더미화
4. R_age와 B_age를 32살을 기준으로 young과 nice_age로 나누고 더미화
5. 연속형 변수에 대해 StandardScaler로 스케일링

Modeling 및 결과_ 모델 설명



1. MLP Classifier

하나 이상의 퍼셉트론(Perceptron)이 여러 개의 층으로 이루어져 있는 구조로, 에러를 최소화시킨다

2. AdaBoost Classifier

Adaptive Boosting으로서 이전 분류기가 틀린 부분을 adaptive하게 바꾸어가며 잘못 분류되는 데이터에 집중하는 방식

3. GradientBoosting Classifier

Gradient가 현재까지 학습된 Classifier의 약점(weak)을 알려주고, 이후 모델이 그것을 중점적으로 보완하는 방식

Modeling 및 결과_ 모델 튜닝

단위 : %



1. MLP

Train = 100
Test = 58.3



Train = 71.9
Test = 64.9

2. AdaBoost

Train = 70.0
Test = 62.2



Train = 68.4
Test = 63.6

3. GradientBoosting

Train = 81.9
Test = 61.9



Train = 76.7
Test = 62.9

Modeling 및 결과_ 모델 비교

단위 : %



분류기	train	test
MLP	71.9	64.9
AdaBoost	68.4	63.6
GradientBoosting	76.7	62.9

Modeling 및 결과_ 모델 성능 검사

단위 : %



분류기	train	test
MLP	71.9	64.9

교차검증 cv=10

분류기	train	test
MLP	74.5	62.8



INDEX 1
UFC / 초기 Data 소개



INDEX 2
EDA / 전처리 및 결측치 처리



INDEX 3
Modeling 및 결과



INDEX 4
한계점

UFC



한계점

한계점



- 실제 경기에선 다양한 변수들이 발생한다.
- 관측 안되는 요소들로 인한 오차들 설명 불가능
ex). 선수 컨디션, 싸움 전략 등등
- 심판의 역량도 아주 무시할 수 없는 부분이다.
- 선수들의 개인 역량이기 때문에 정확한 수치로 나타낼 때 나타나는 오차 설명 불가능
- 짧은 시간내에도 변화가 큰 스포츠 세계를 반영하지 못한 데이터
: 스포츠는 1년 내에도 큰 변화가 있는 분야인데 다음의 데이터는 1998년 이후의 누적 데이터로 큰 변화를 무력화 시키는 경향이 있다.
: lag를 조정 후 나온 결측치를 처리 시 데이터손실이 더 크다고 판단

