UFC Historical Match data

OSEMN pipeline

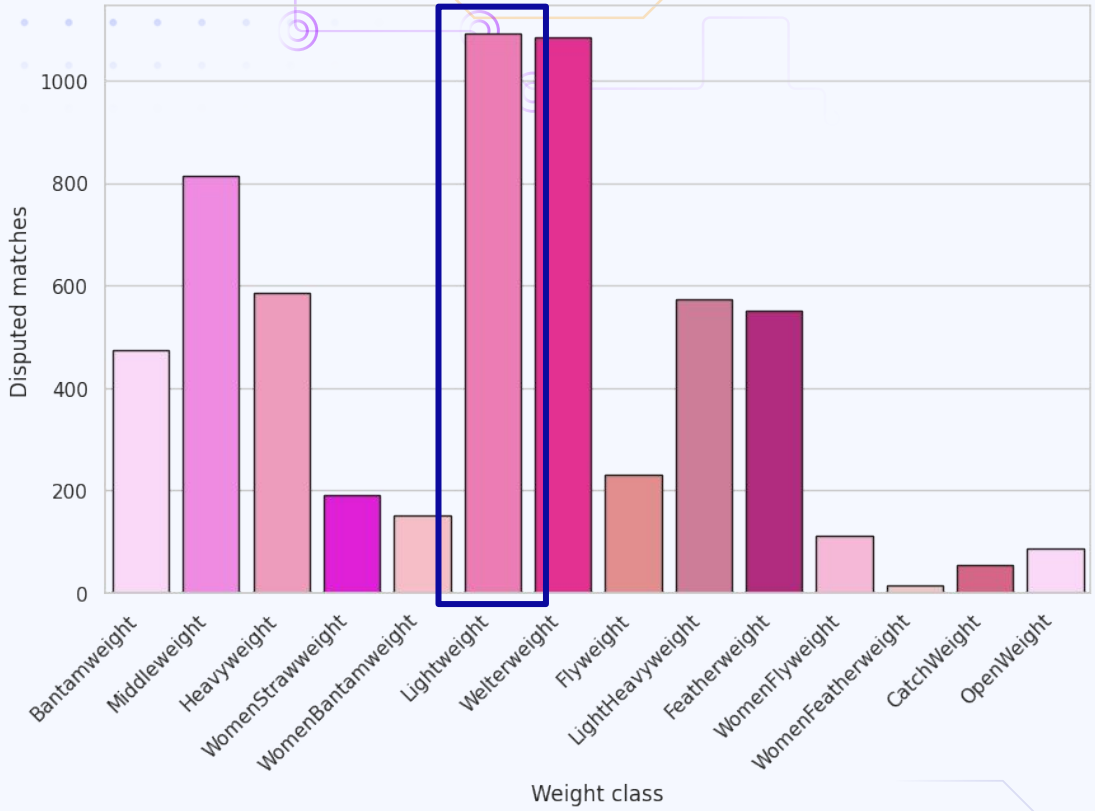
LUISS ‘Guido Carli’ - MSc in Data Science and Management
Python and R Course, Final Project - A.Y. 2023-24
Group 5: Coci Marco - Cecere Rachele - Isayas Aida - Marchioni Gian Lorenzo

OBTAIN

- The dataset contains information of 6012 pros, it also contains 144 columns
- We reduced it and picked only certain features based on our exploration
- Not a tidy format yet
    - Most are duplicates since features for Red and Blue fighters are listed separately

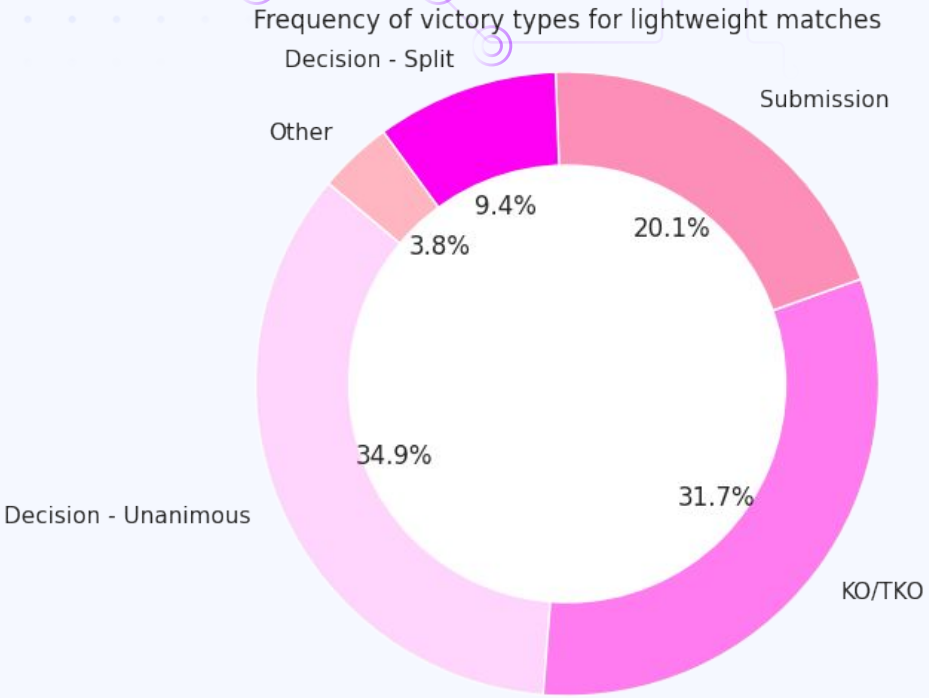
	R_fighter	B_fighter	Referee	date	location	Winner	title_bout	weight_class	B_avg_KD	B_avg_opp_KD	...	R_win_by_Decision_Unanimous	R_win_by_I
0	Adrian Yanez	Gustavo Lopez	Chris Tognoni	2021-03-20	Las Vegas, Nevada, USA	Red	False	Bantamweight	0.000	0.0	...	0	
1	Trevin Giles	Roman Dolidze	Herb Dean	2021-03-20	Las Vegas, Nevada, USA	Red	False	Middleweight	0.500	0.0	...	0	
2	Tai Tuivasa	Harry Hunsucker	Herb Dean	2021-03-20	Las Vegas, Nevada, USA	Red	False	Heavyweight	NaN	NaN	...	1	
3	Cheyanne Buys	Montserrat Conejo	Mark Smith	2021-03-20	Las Vegas, Nevada, USA	Blue	False	WomenStrawweight	NaN	NaN	...	0	
4	Marion Reneau	Macy Chiasson	Mark Smith	2021-03-20	Las Vegas, Nevada, USA	Blue	False	WomenBantamweight	0.125	0.0	...	1	

FIGHTERS WEIGHT CLASS DISTRIBUTION



- We focus on Lightweight class: highest count data and is the most popular and known weight class
- Welterweight is the most frequent weight class type after Lightweight

FEATURES SELECTION FROM WIN TYPE



- We took an additional data set related to ours from the same source
- We want to see what could be good predictors
- Fighters win via **KO/TKO** or **Submission** or **Decision** and we took features related later

SCRUBBING

Variables that **we selected** that most **likely influence** victories:

1. Physical characteristics

2. Performance indicators:
   a. Avg. Significant strikes
   b. Avg. Head strikes
   c. Avg. Takedowns
   d. Avg. Knockdowns

3. Victory and experience statistics:
   a. Amount of wins and losses
   b. Streaks of victories and defeats
   c. Amount of rounds fought

	count	mean	std	min	25%	50%	75%	max
B_total_rounds_fought	1091.0	10.141155	13.004454	0.00	1.000000	6.000000	14.000000	86.000000
B_current_win_streak	1091.0	0.841430	1.308956	0.00	0.000000	0.000000	1.000000	7.000000
B_current_lose_streak	1091.0	0.430797	0.676964	0.00	0.000000	0.000000	1.000000	5.000000
B_wins	1091.0	2.692942	3.604551	0.00	0.000000	1.000000	4.000000	23.000000
B_longest_win_streak	1091.0	1.626031	1.760270	0.00	0.000000	1.000000	3.000000	8.000000
B_losses	1091.0	1.656279	2.110721	0.00	0.000000	1.000000	2.000000	14.000000
B_Height_cms	1087.0	176.746145	4.891668	160.02	172.720000	177.800000	180.340000	193.040000
B_Reach_cms	932.0	181.471009	5.593063	162.56	177.800000	180.340000	185.420000	203.200000
B_age	1062.0	28.625235	3.673894	18.00	26.000000	28.500000	31.000000	40.000000
B_avg_SIG_STR_landed	828.0	32.652425	20.632936	0.00	18.000000	30.000000	43.375000	131.739330
B_avg_HEAD_landed	828.0	20.538878	14.614758	0.00	10.546875	17.766541	27.054688	111.637010
B_avg_KD	828.0	0.209219	0.330221	0.00	0.000000	0.002443	0.312500	2.373047
B_avg_TD_landed	828.0	1.215497	1.373016	0.00	0.125000	0.962524	1.779785	10.875000
R_total_rounds_fought	1091.0	15.194317	14.800595	0.00	4.000000	11.000000	22.000000	77.000000
R_current_win_streak	1091.0	1.229148	1.564137	0.00	0.000000	1.000000	2.000000	12.000000
R_current_lose_streak	1091.0	0.416132	0.647561	0.00	0.000000	0.000000	1.000000	4.000000
R_wins	1091.0	4.098992	4.029628	0.00	1.000000	3.000000	6.000000	23.000000
R_longest_win_streak	1091.0	2.368469	2.031377	0.00	1.000000	2.000000	3.000000	13.000000
R_losses	1091.0	2.345555	2.483104	0.00	1.000000	2.000000	3.000000	14.000000
R_Height_cms	1090.0	176.623211	4.847889	165.10	172.720000	175.260000	180.340000	193.040000
R_Reach_cms	1044.0	181.288851	5.617913	165.10	177.800000	180.340000	185.420000	203.200000
R_age	1085.0	28.750230	3.813855	19.00	26.000000	29.000000	31.000000	41.000000
R_avg_SIG_STR_landed	982.0	34.387591	21.602975	0.00	20.000000	29.936401	44.240486	212.142956
R_avg_HEAD_landed	982.0	21.939575	15.192424	0.00	12.000000	18.950045	28.046875	173.163765
R_avg_KD	982.0	0.208317	0.308753	0.00	0.000000	0.033752	0.312500	2.000000
R_avg_TD_landed	982.0	1.331252	1.336092	0.00	0.252382	1.000000	2.000000	9.000000

MISSING DATA MATRIX

3 types of Null Values in our data set:

- Missing Completely at Random (MCAR)like: *Height_cms*
- Missing at Random (MAR) like: *Reach_cms*
- Not Missing at Random (MNAR)

```
Missing Values Summary:
R_fighter                  0
B_fighter                  0
Winner                     0
weight_class               0
date                       0
B_total_rounds_fought      0
B_current_win_streak       0
B_current_lose_streak      0
B_wins                     0
B_longest_win_streak       0
B_losses                   0
B_Height_cms               4
B_Reach_cms              159
B_age                     29
B_avg_SIG_STR_landed     263
B_avg_HEAD_landed        263
B_avg_KD                 263
B_avg_TD_landed          263
R_total_rounds_fought      0
R_current_win_streak       0
R_current_lose_streak      0
R_wins                     0
R_longest_win_streak       0
R_losses                   0
R_Height_cms               1
R_Reach_cms               47
R_age                      6
R_avg_SIG_STR_landed     109
R_avg_HEAD_landed        109
R_avg_KD                 109
R_avg_TD_landed          109
dtype: int64
```

MANAGE NA VALUES

MNAR: missingness is directly related to the nature of the data itself -> new fighters won't have historical data

POSSIBLE SOLUTION:
Take the minimum value for all future fights of an athlete.

ISSUES: When we looked at the results, we realized that the median shifted a lot downwards and would alter our results.

	count	mean	std	min	25%	50%	75%	max
B_total_rounds_fought	735.0	14.457143	13.665845	1.00	5.000000	10.000000	19.000000	86.000000
B_current_win_streak	735.0	1.213605	1.435946	0.00	0.000000	1.000000	2.000000	7.000000
B_current_lose_streak	735.0	0.529252	0.739247	0.00	0.000000	0.000000	1.000000	5.000000
B_wins	735.0	3.874830	3.806721	0.00	1.000000	3.000000	5.000000	23.000000
B_longest_win_streak	735.0	2.318367	1.692869	0.00	1.000000	2.000000	3.000000	8.000000
B_losses	735.0	2.308844	2.252728	0.00	1.000000	2.000000	3.000000	14.000000
B_Height_cms	735.0	176.825469	4.583179	165.10	172.720000	177.800000	177.800000	193.040000
B_Reach_cms	735.0	181.521878	5.520192	167.64	177.800000	180.340000	185.420000	203.200000
B_age	735.0	29.066667	3.549136	20.00	27.000000	29.000000	31.000000	40.000000
B_avg_SIG_STR_landed	735.0	34.323512	20.457696	0.00	19.500000	31.738525	45.221578	131.739330
B_avg_HEAD_landed	735.0	21.709324	14.710382	0.00	11.718750	18.994141	28.532227	111.637010
B_avg_KD	735.0	0.224836	0.337139	0.00	0.000000	0.023438	0.400542	2.373047
B_avg_TD_landed	735.0	1.245276	1.389055	0.00	0.226562	1.000000	1.781250	10.875000
R_total_rounds_fought	735.0	19.103401	15.169620	1.00	7.000000	15.000000	27.000000	77.000000
R_current_win_streak	735.0	1.542857	1.693231	0.00	0.000000	1.000000	2.000000	12.000000
R_current_lose_streak	735.0	0.423129	0.661728	0.00	0.000000	0.000000	1.000000	4.000000
R_wins	735.0	5.227211	4.094100	0.00	2.000000	4.000000	8.000000	23.000000
R_longest_win_streak	735.0	2.967347	2.016018	0.00	2.000000	3.000000	4.000000	13.000000
R_losses	735.0	2.848980	2.633425	0.00	1.000000	2.000000	4.000000	14.000000
R_Height_cms	735.0	176.479891	4.622954	165.10	172.720000	175.260000	177.800000	190.500000
R_Reach_cms	735.0	181.297252	5.524761	167.64	177.800000	180.340000	185.420000	195.580000
R_age	735.0	29.352381	3.745248	19.00	27.000000	29.000000	32.000000	41.000000
R_avg_SIG_STR_landed	735.0	36.116811	21.987430	0.00	21.296875	31.302986	46.072266	212.142956
R_avg_HEAD_landed	735.0	23.049070	15.637594	0.00	12.675781	19.839844	28.787476	173.163765
R_avg_KD	735.0	0.216521	0.299450	0.00	0.000000	0.062500	0.371094	1.625000
R_avg_TD_landed	735.0	1.321765	1.285947	0.00	0.301865	1.000000	2.000000	7.000000

TIDYING

To use the **'hue'** argument of the plot functions:

- We need to **reshape the data frame into a tidy format:**
  - The same column of stats for each fighter will be concatenated and stored in a single column
  - Add a corner color column

	Rounds fought	Current win streak	Longest win streak	Victories	Avg Head Strikes	Avg Significant Strikes	Avg knockdowns	Avg takedowns	Reach (cm)	Height (cm)	Ape index (cm)	Age	Corner
0	18	1	6	7	32.281250	42.18750	0.31250	0.632812	190.50	182.88	7.62	41.0	Red
1	15	0	3	4	58.656250	67.87500	0.25000	0.031250	182.88	177.80	5.08	25.0	Red
2	17	1	6	7	20.085938	27.87500	0.56250	1.367188	177.80	177.80	0.00	29.0	Red
3	0	0	0	0	NaN	NaN	NaN	NaN	180.34	185.42	-5.08	27.0	Red
4	12	2	2	4	16.781250	30.65625	0.53125	0.875000	182.88	175.26	7.62	28.0	Red

Hue: The term "hue" refers to the attribute of a color that distinguishes it from other colors, such as red, blue, or green

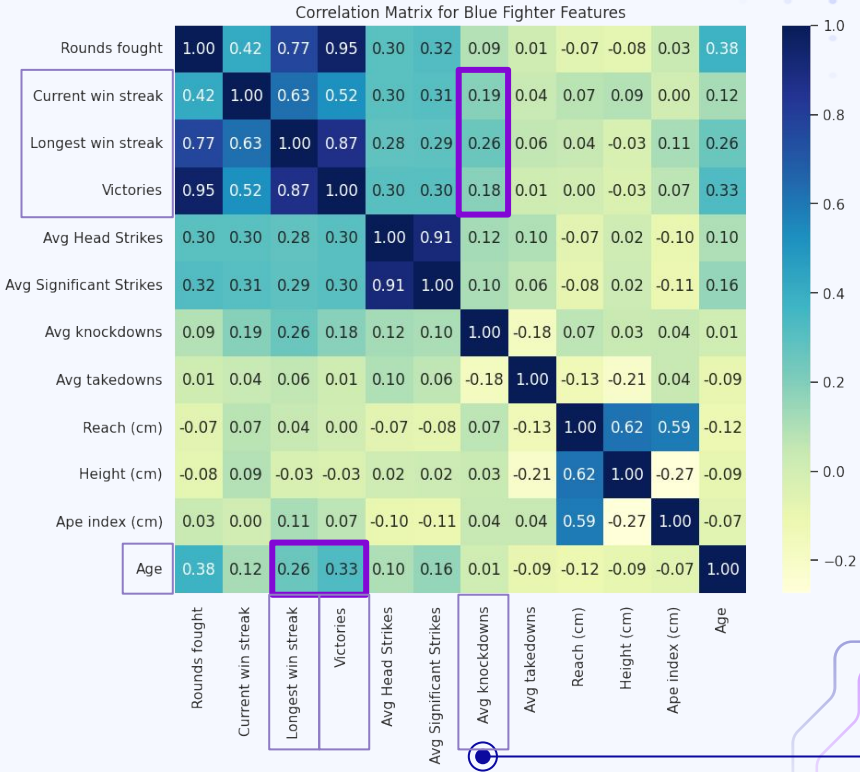
EXPLORATION - HEATMAPS

- Combined generic heatmap for red and blue corner
- Difficult to identify any significant relationship

HEATMAPS SEPARATED BY COLOR

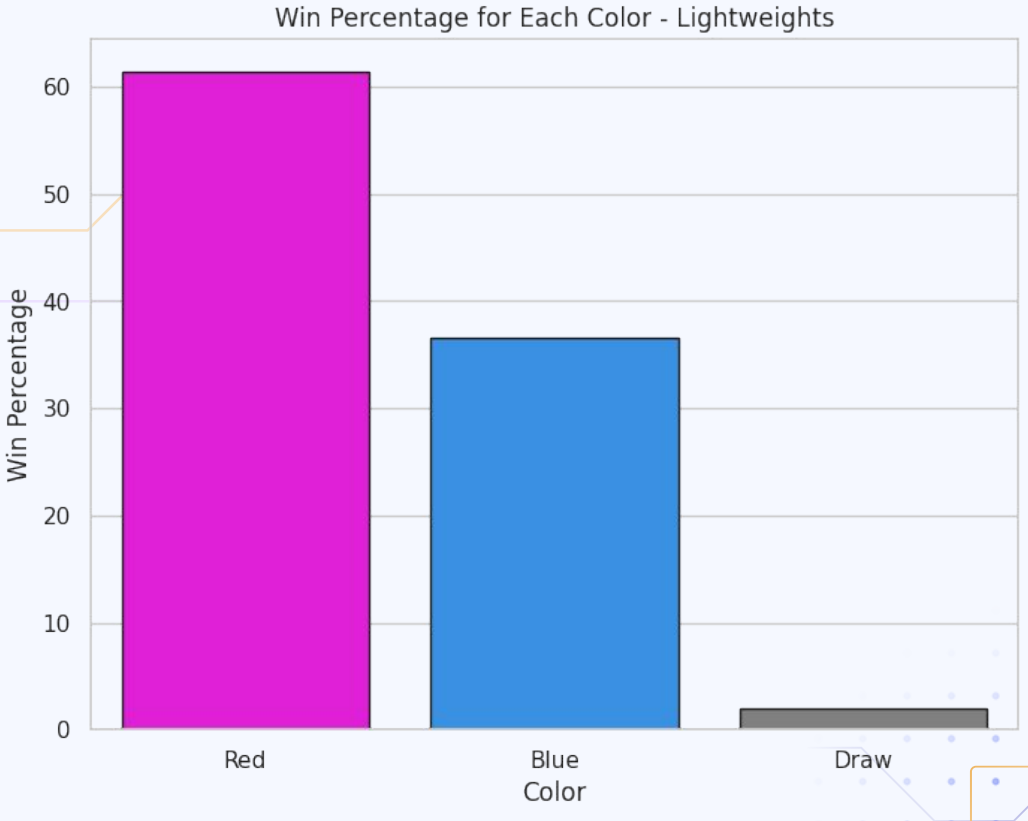
Red Corner

Blue Corner

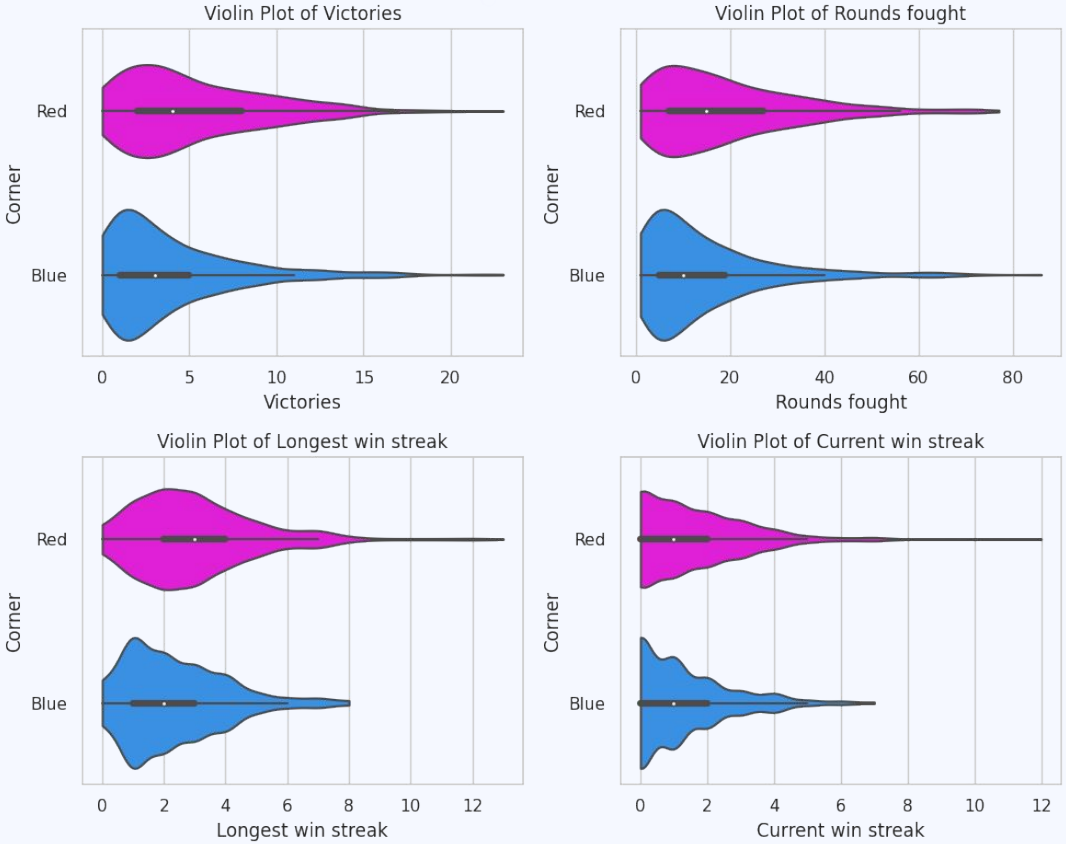


CORNER COLOR BIAS

Fighters in the **Red Corner** win significantly more



FURTHER INVESTIGATION ON COLOR BIAS



On average, the **Red Corner** is assigned to UFC athletes with **more experience** in terms of rounds fought and match victories.

PREPROCESSING and MODELLING

Model selection

- Correlation matrices do not suggest evident correlations between quantitative features
- Victory of a match is a binary response variable
- The model of choice is then **logistic regression**

Preprocessing

It is **evident** from the EDA that **color is related to other features**

- Separate it from fighters' names
- Perform another logistic regression to quantify relation with other features
- Use it as a predictor in final model

PREPROCESSING AND MODELLING

Fighter anonymization

- Sample a color randomly, assign label ‘ Fighter 1’ (F1) to that corner.
- The response variable becomes **‘F1 victory’**
- ‘F1 color’ becomes a feature
- Others are the difference between F1 and F2 features:
  e.g. ‘Height of F1 minus height of F2’

Splitting and scaling

- Split final 735 entries into 80-20 training and test sets
- Compute differences
- Standard scaling using ‘`preProcess`’ from ‘`Caret`’ library

MODEL TRAINING AND PERFORMANCE

Backwards selection

ASI backwards selection algorithm starting from 12 features.

After selection:

Height difference	**TD difference**
Age difference	**Difference in victories**
Difference in number of rounds	**Color**

Cross-validation

- 5-fold cross validation

  Test accuracy: 65-68%
  Informedness: 0.25

- Predictions are better when including all suggested predictors even if related: priority to prediction

CONCLUSIONS

Fighter color is a good predictor in the model:

- Confirms findings from scrubbing and EDA: color assignment is not random
- Possible improvements: explore different or more extensive feature selection techniques, incorporate interaction terms

The **model outperforms the trivial classifier** and shows **no significant overfitting**.

THANKS FOR YOUR ATTENTION

QUESTIONS?