Sportsbook&Casino Betting on UFC fights

1st January 2020

OVERVIEW

Recently S&C, is a company that started doing gambling on sports, and they want to invest on Mixed martial arts. They want exactly to work with UFC. And in order to maximize their benefice.They hire you as a data scientist to gather data and make statistical analysis on them, and machine learning to predict the winner based on historical data.

DATA

To gather your data it's always good to visit websites that have historical data, and if it is legal, you can webscrap the data and clean it.

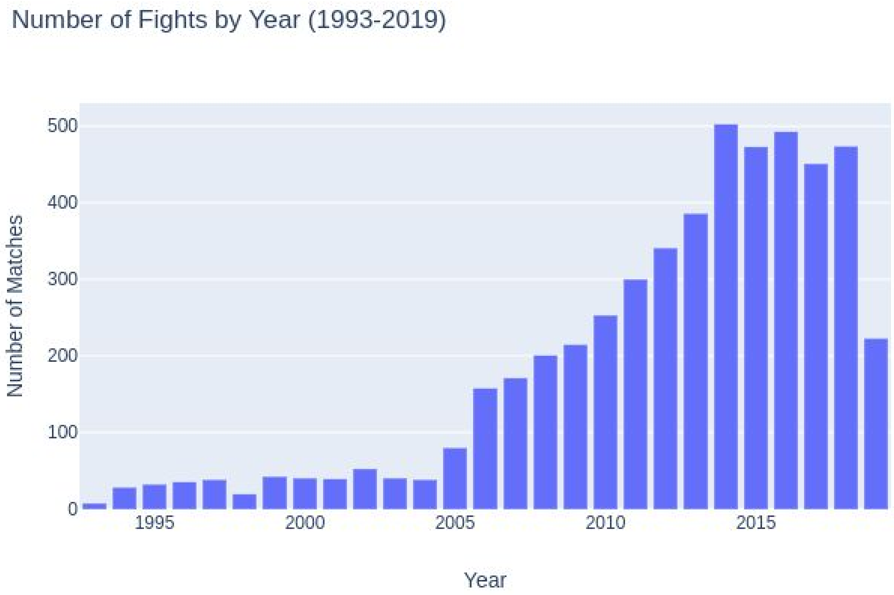
Fortunately, you found a good dataset on kaggle named "UFC Fights", that contains data on fights since 1993 to 2019, and all fighters data. Still, it needs to be cleaned.

In the dataset, it exists columns starting with `R_`/`B_ refering the red or blue corner.` Fortunately KD/TD number of kickdown/Takedown. Win_by is the method in which the winner won, it can be KO (Knocked-out), TKO (the medicine or the referee ended the fight because he saw the fighter can't continue), DECISION (Jury decision after finishing the rounds). Last_round and last_round_time are the round in which the fight ended and in which time. There exists also head and body referring to stikes to the body or the head during the match. Concerning the fighter, we have his height, reach, age, wins /draws (how many times he wins/draws he has in his ufc career), alongside with how he wins in columns starts with win_by and ends with what type (ex: submissions in win_by_Submissions column).

DATA EXPLORATION

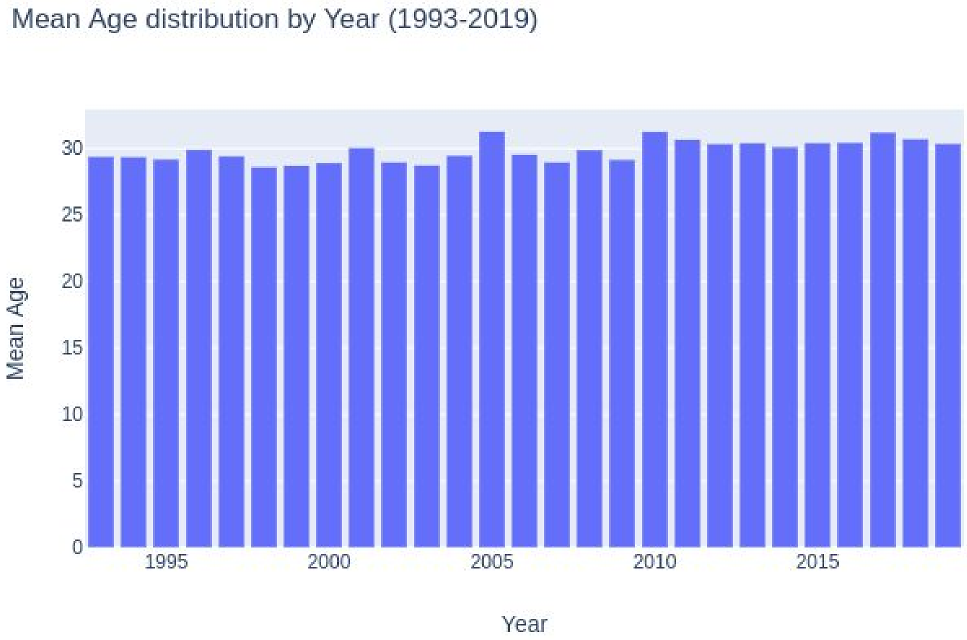
Let’s see what’s in our data set. You can always see how I cleaned this data set on this notebook.

UFC every year organises many events:



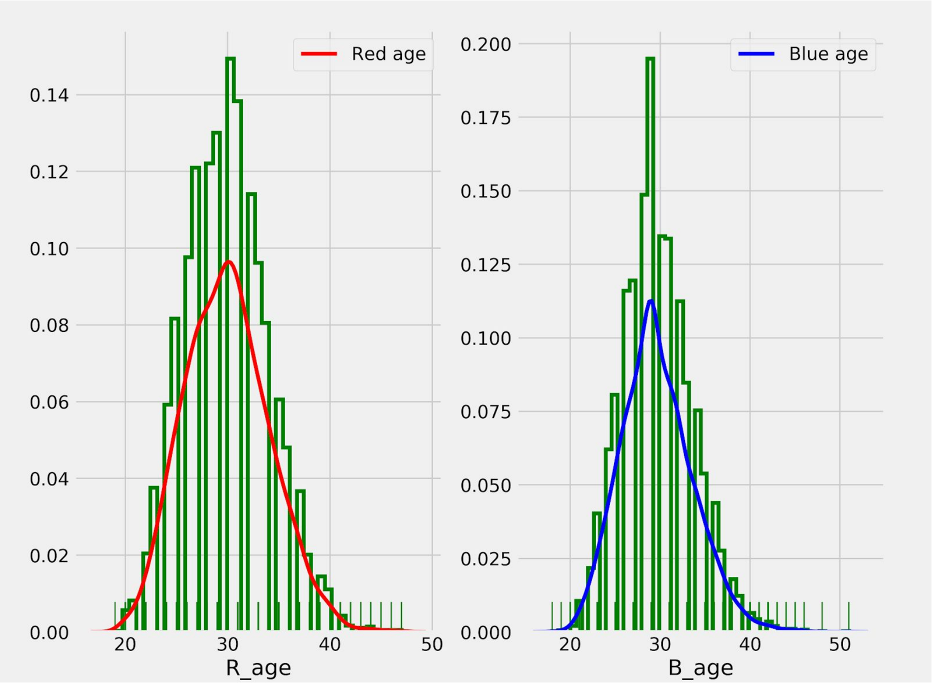
In 2014 it has more than 500 events.

Fighters mean age by years should be almost constant, but in this graph:



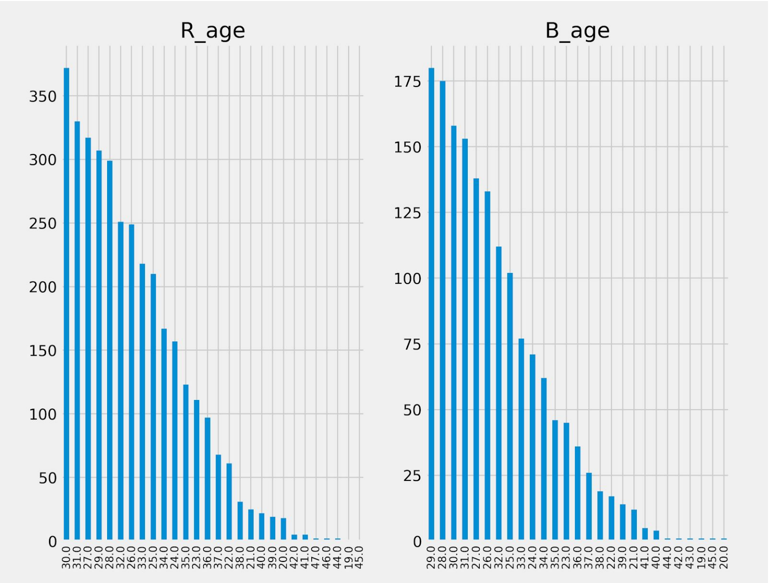
We remark some differences but not too much, this due to some fighters age like: Ron van Clief, who has his last fight at the age of 51, or like Sean Daugherty who started his UFC career at 21.

This graph give more insights on fighters age:



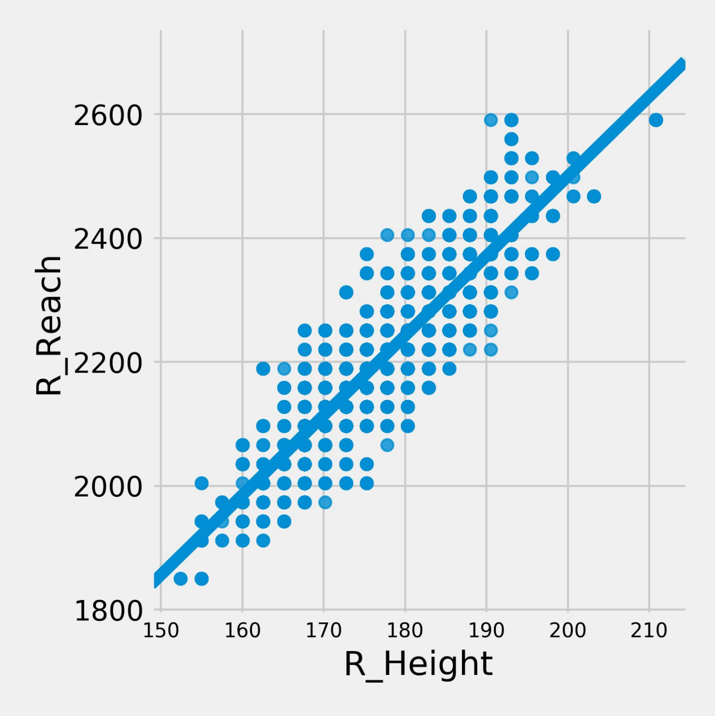
Clearly, many fighters ages are 30, because at this age the fighter gain experience and power.

Let’s see winner’s age:

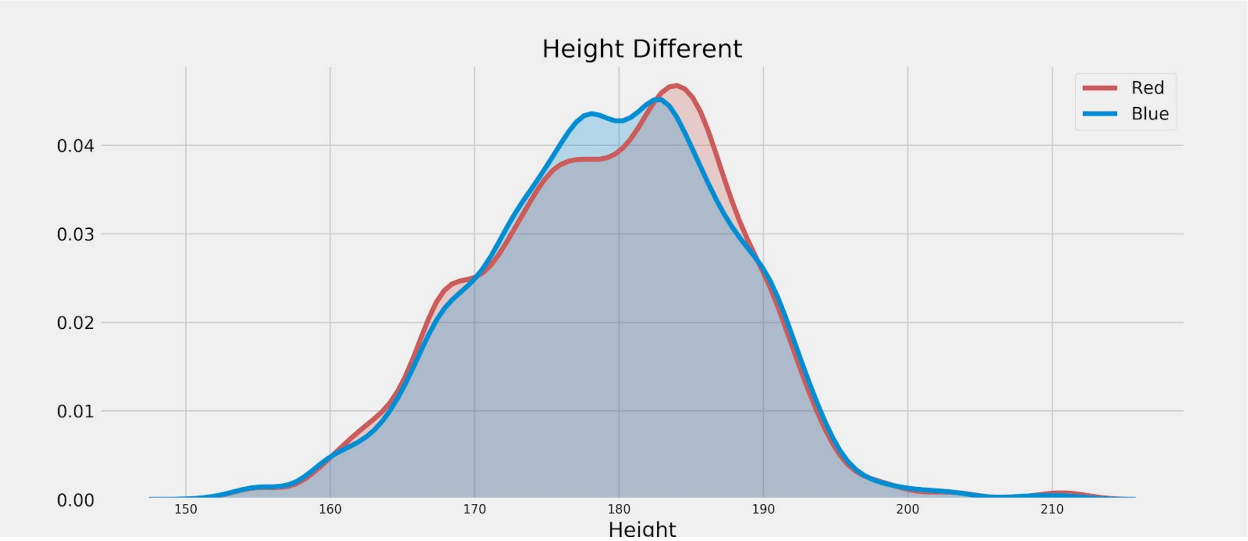


It’s true age has a big effect on fight winning. If we consider age only, our oldest and youngest fighters have a lower probability in winning on a 30 years old fighter.

Height and Reach of the fighter have a linear relation, and this is logical because the taller the height of human the taller goes his whole body, hence long reach.

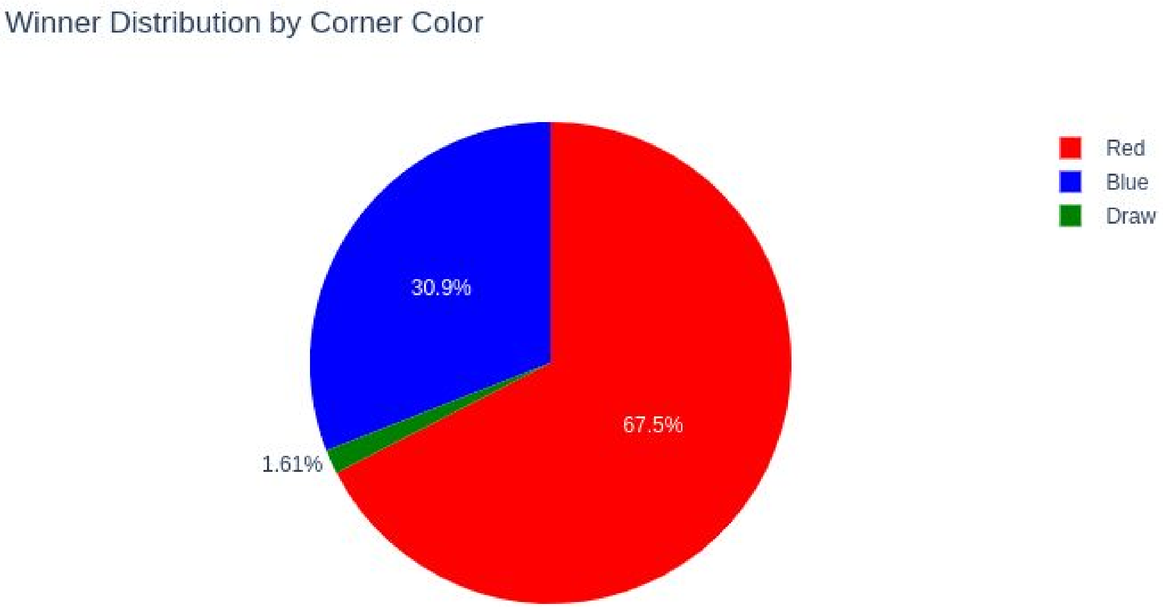


Does the fighters in the ring have the same height?



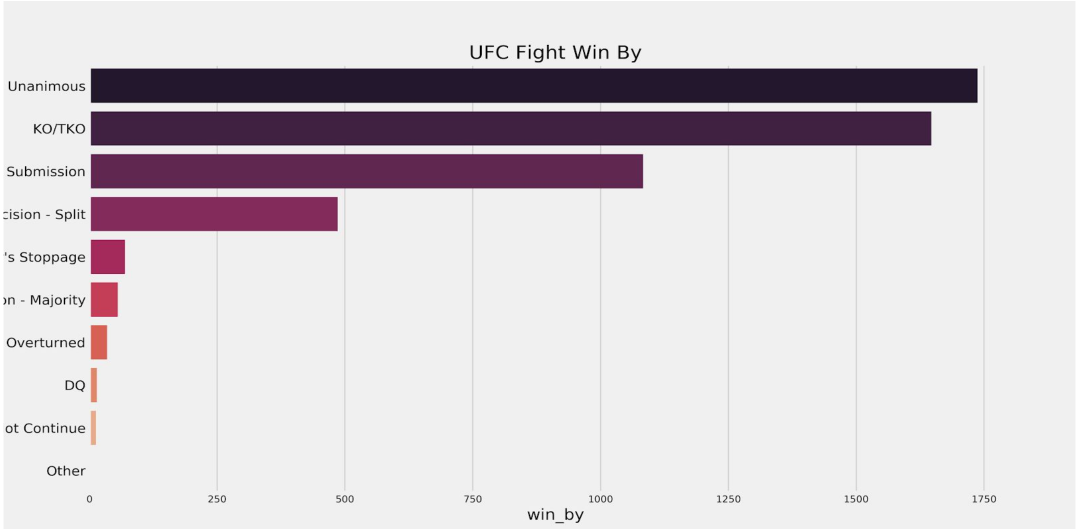
This is because fights are organized in weight class, and not every tall fighter is going to be Heavyweight. Still, this will have an effect on the winning, the taller the fighter, the long gonna be the reach, Hence more punches.

Red corner and Blue corner, let’s see who wins more:



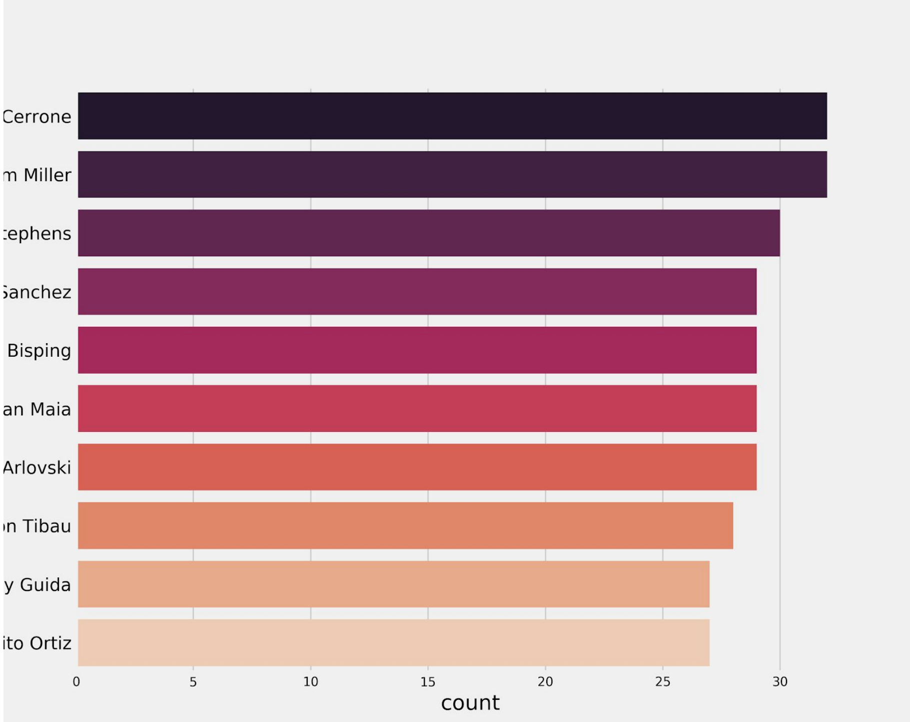
Interesting! The corner of the fighter has a big effect on the winning.

Let’s see how those fights end:



Many fights ended by jury decision, followed by KO/TKO. Not much difference though. Most fight ended with Jury decision because most of fights had a fixed number of rounds.

What are some fighters who had many fights in UFC:



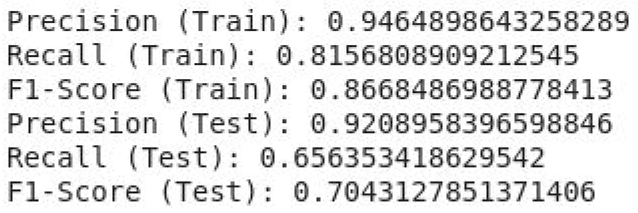
MACHINE LEARNING

Our dataset has many categorical columns. We want to make a prediction based on those columns. The prediction (target) is the winner.

We use for this problem CatBoost. CatBoost is a high-performance open source library for gradient boosting on decision trees. It Improves your training results, that allows you to use non-numeric factors, instead of having to pre-process your data or spend time and effort turning it to numbers.

I had a good result with this model and it may be better if there will be some adjustment on the data.

My results:



```
Precision (Train): 0.9464898643258289
Recall (Train): 0.8156808909212545
F1-Score (Train): 0.8668486988778413
Precision (Test): 0.9208958396598846
Recall (Test): 0.656353418629542
F1-Score (Test): 0.7043127851371406
```