Classification problem : predict the result of a UFC match (in terms of probabilities)

Context

The UFC is nowadays the biggest mixed martial arts competition organization in the world in terms of views and prestige. The fighters fighting in this one are usually considered as the best in the world (they often came from different organisations and got promoted there after a long road). People have always wanted to predict the winner of those kind of fights and even more today as bettings on such games are now becoming so huge that it is becoming closer to what a boxing match could make for sponsors and sport bets organisations concerning earnings. To be able to predict (or at least try at best) the final result of a UFC match (dropping every stats of the concerning fight in the model and only keeping stats up until the actual fight) could be a good betting decision helper and therefore making fights even more exciting to watch.

Description of the dataset

This dataset is a list of every UFC fight in the history of the organisation. Every row contains information about both fighters, fight details and the winner. The data was scraped from ufcstats website. The link towards the dataset can be found here : https://www.kaggle.com/rajeevw/ufcdata.

Each row is a compilation of both fighter stats. Fighters are represented by 'red' and 'blue' (for red and blue corner). So for instance, red fighter has the complied average stats of all the fights except the current one. The stats include damage done by the red fighter on the opponent and the damage done by the opponent on the fighter (represented by 'opp' in the columns) in all fights this particular red fighter has had, except this one as it has not occured yet (in the data). Same information exists for blue fighter. The target variable is 'Winner' which is the only column that tells you what happened. Here are some column definitions :

- `R_` and `B_` prefix signifies red and blue corner fighter stats respectively
- `_opp_` containing columns is the average of damage done by the opponent on the fighter
- `KD` is number of knockdowns
- `SIG_STR` is no. of significant strikes 'landed of attempted'
- `SIG_STR_pct` is significant strikes percentage
- `TOTAL_STR` is total strikes 'landed of attempted'
- `TD` is no. of takedowns
- `TD_pct` is takedown percentages
- `SUB_ATT` is no. of submission attempts
- `PASS` is no. times the guard was passed?
- `REV` ???
- `HEAD` is no. of significant strikes to the head 'landed of attempted'
- `BODY` is no. of significant strikes to the body 'landed of attempted'
- `CLINCH` is no. of significant strikes in the clinch 'landed of attempted'
- `GROUND` is no. of significant strikes on the ground 'landed of attempted'
- `win_by` is method of win
- `last_round` is last round of the fight (ex. if it was a KO in 1st, then this will be 1)
- `last_round_time` is when the fight ended in the last round
- `Format` is the format of the fight (3 rounds, 5 rounds etc.)
- `Referee` is the name of the Ref

- `date` is the date of the fight
- `location` is the location in which the event took place
- `Fight_type` is which weight class and whether it's a title bout or not
- `Winner` is the winner of the fight
- `Stance` is the stance of the fighter (orthodox, southpaw, etc.)
- `Height_cms` is the height in centimeter
- `Reach_cms` is the reach of the fighter (arm span) in centimeter
- `Weight_lbs` is the weight of the fighter in pounds (lbs)
- `age` is the age of the fighter
- `title_bout` Boolean value of whether it is title fight or not
- `weight_class` is which weight class the fight is in (Bantamweight, heavyweight, Women's flyweight, etc.)
- `no_of_rounds` is the number of rounds the fight was scheduled for
- `current_lose_streak` is the count of current concurrent losses of the fighter
- `current_win_streak` is the count of current concurrent wins of the fighter
- `draw` is the number of draws in the fighter's ufc career
- `wins` is the number of wins in the fighter's ufc career
- `losses` is the number of losses in the fighter's ufc career
- `total_rounds_fought` is the average of total rounds fought by the fighter
- `total_time_fought(seconds)` is the count of total time spent fighting in seconds
- `total_title_bouts` is the total number of title bouts taken part in by the fighter
- `win_by_Decision_Majority` is the number of wins by majority judges decision in the fighter's ufc career
- `win_by_Decision_Split` is the number of wins by split judges decision in the fighter's ufc career
- `win_by_Decision_Unanimous` is the number of wins by unanimous judges decision in the fighter's ufc career
- `win_by_KO/TKO` is the number of wins by knockout in the fighter's ufc career
- `win_by_Submission` is the number of wins by submission in the fighter's ufc career
- `win_by_TKO_Doctor_Stoppage` is the number of wins by doctor stoppage in the fighter's ufc career

Plan

The problem is to build a strong model based on fighters stats until the fight we want to predict (meaning that the model should be updated after each UFC event so the last stats get updated and then the model remains up-to-date) that is efficient in predicting the winner of each fight during an event.

The key steps to complete on this project development are the following :

- Data cleaning
- Quick Exploratory Data Analysis to have a better knowledge on our variables
- Feature engineering
- Preprocessing, split data (prepare data to feed our models)
- Build several classification models, optimize them and compare their accuracies difference (precision, recall)