Ultimate Fighting Championship's Fights Analysis

Name: Binh Vu, ID: 394, Email: Binhhtvu@gatech.edu

Date: Nov 27, 2022, Course: ISYE 7406

Abstract:

As combat sports increase in popularity, causal viewers often overlook the complexity of mixed martial art (MMA). In the early day of the UFC, when fighters' skills are more one-dimensional (wrestling only or boxing only), one may have a higher chance of predicting a winner based on the fighting style. However, to compete in today's competition, successful fighters often have a multifarious understanding of all MMA which directly contributes to the complexity of about. In this paper, we will attempt to predict the winner of a fight based on historical fight data combined with the fighter's records and the fighter's physical attributes. Along the way, we will explore the advantage of physical attributes and try to find the best machine-learning model to predict the winner. We will first perform some feature engineering on the given dataset. Then we will use the random forest as a baseline model to compare logistic regression (LR), linear discriminant analysis (LDA), ridge regression, and boosting. After, we will utilize some optimization technics such as standardization, cross-validation, and feature selections to attempts to find a better accuracy score. As it turns out, you can predict a fight with a 66.7% accuracy score using logistic regression with only data engineering. Additionally, based on the dataset, a positive correlation of fight winners is drawn based a clear advantage with reach/age. This finding is interesting because it suggests a cycle in a combat sport where skills advantage beats physical attributes, but physical attributes play a bigger role when both fighters are equally skilled. In other words, you can beat a bigger man with skills but not if that bigger man also has the same thing.

Introduction:

Combat sports have increased in popularity in recent years thanks to covid lockdown and the shutting down of most sports. Fighting has been part of humanity since the very beginning as we fought wild animals in the African plane, fought against other archaic humans, and eventually fought among ourselves. Throughout the infighting, various groups/cultures form a distinctive fighting style that leads to a diversity of fighting technics or commonly known as a martial art. Different martial art have a distinct advantage over others (wrestler vs boxer) or similarly fighting stances (southpaw vs orthodox) and have different strengths and weaknesses. The Ultimate Fighting Championship or UFC is an organization that brings together different mixed martial arts under a single unified rule set. Fighters compete in 1 vs 1 combat inside an octagon-shaped cage with a referee to stop the fight as need be based on the specified safety rules and standards. Combat sport is a complex game that requires the fighter to have a minimum physical and mental toughness. UFC fighters are considered by most to be the best of the best. As the two fighters' skills increase, the competition becomes more of a chess match with each fighter attempting to anticipate the opponent's next move and strategize a countermeasure. Even if the fight seems predictable, a fighter can always have the element of surprise. Take UFC 270: Francis Ngannou (heavy weight champion) vs Ciryl Gane Interim Champion) for example. Francis, regarded as one of the most powerful punchers in the heavyweight division, has resorted to using his wrestling skills to beat Ciryl Gane known for his evasive footwork. Most analysts never predicted that Francis Ngannou would beat Ciryl Gane in round 5 with wrestling but that is part of the unpredictability of combat sport.

Problem Description

Given the complexity of combat sports and all its unpredictability, it would be improbable to quantify all the predictors of a fighter. As such, we are using some historical fight stats combined with fight odds and fighters' stats to predict a winner of a fight. We will first perform some data cleaning which includes combining some predictors that are closely related and filling in missing values. Then, we will train 5 machine learning models to predict the dataset. Afterward, we will try to optimize the accuracy score through data scaling, feature selection, and cross-validation. Lastly, we will collect the accuracy score for each iteration and compare the accuracy score to determine the best model.

Problem Statement

The goal of my research paper is to answer the following questions:

1. Does physical attributes have an impact on the chance of winning?
2. Is betting odds a good indicator of a fighter's chance of winning?
3. Can we predict which fighter is going to win with high degrees of accuracy?
4. What is the best ML model with the highest accuracy?

Data Set Sources:

Aggregate of three separate data source: on Kaggle.

1. ufcstats.com - (bout and fighter statistics)
2. Bestfightodds.com - (odds)
3. Kagge.com/martj42/ufc-rankings - (ufc rankings)

Since the data came from three separate sources, there will inevitably be some missing values. The list of the top 7 predictors with missing values is shown in table 1. A quick inspection seems to indicate that most of the missing values come from the ranking of the various women's rankings. Since the early day of the UFC, male fighters have always outnumbered female fighters (Fig 1). Since the ranking only counts the top 15 fighters in a division that may have hundreds of fighters, that reason explained why the women's weight ranking contains a lot of null values. In fact, all ranking predictors suffer from a similar problem. Another problem with this dataset is the separation of the similar predictor. For instance, 'R_Height_cms' & 'B_Height_cms' can be combined into one single value. Lastly, predictors contain different scaling. This can be seen with 'no_of_round' with values ranges from [3-5] while 'total_fight_time_secs' range [5,1500].

Table 1: Top 7 null Values Variable

Predictors Name	Num of Null
B_Women's Featherweight_rank	4896
R_Women's Featherweight_rank	4889
B_Pound-for-Pound_rank	4861
B_Women's Flyweight_rank	4852
R_Women's Flyweight_rank	4837
B_Women's Strawweight_rank	4835
B_Women's Bantamweight_rank	4818

Data Analysis:

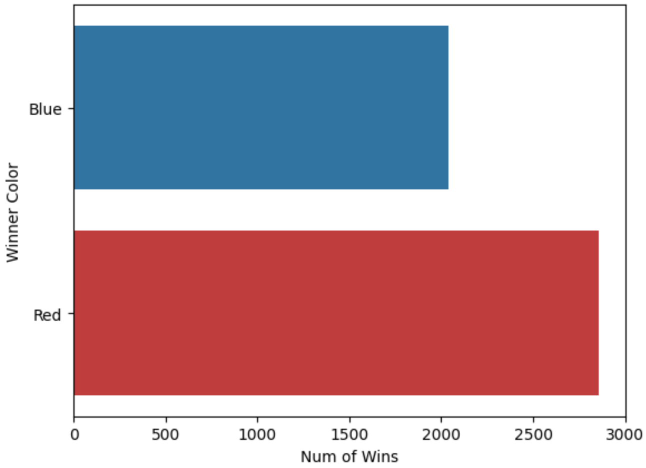
Data Set Overview

The data contains 4896 raw fight data with a combined 119 attributes about the fight odds, fighters' attributes (height, weight, reach, etc.), fighters' fight history (win/lose), and ranking. Full list of predictors is included in appendix.

The notable column is the winner which has categorical values of red or blue and no missing value. Red and blue are the designation of a fighter which changes with each fight. Overall, 67 predictors out of 119 were without missing values while other predictors have varying degrees of missing values.

Data Set Visual Exploration

Out of curiosity, we wanted to see the distribution between the blue and red winners. we were expecting the distribution between the two to be equal since the color designation is randomized. However, it was a surprise that the red fighter has noticeable wins over the opponent. A possible explanation for this could be because of the psychological underlying representation of each color; red is associated with aggression/attack and blue is defensive. Aggressive fighters don't



necessarily mean you have a higher chance of winning but throwing more punches or offensive wrestling could possibly give you more chances to win the match. However, this explanation requires the underlying psychological explanation to hold true all the time, which it is not.

Most fight ends in round 3 (Fig 2). This makes sense because most fights are planned for 3 rounds and even 5 round fights could end in 3 rounds (Fig 3). The first-round finish is the next highest. This also explain the 'Connor McGregor' of the fight game; High power/ strong punching power during the early round but drop off rapidly in successive rounds.

Most finishes are through Unanimous Decision (U-DEC) with Knock-Out (KO)/ Technical-Knock-Out (TKO) and Submission (SUB) following (Fig 4). This means that most of the UFC fight ended with a clear winner with some Split-Decision (S-DEC). Specifically, most fights that do not go to a decision round end in a punch, punches, or rear naked choke (Fig 5). The rarest finishes are Ankle Locks, Key Lock, and Peruvian Necktie all of which are a type of submission technics.

There is a saying in gambling - 'The house always wins.' That means the gambling establishment will always have the advantage. The two predictors that could prove to be useful when it comes to predicting the winner are 'R_ev'/'B_ev' and 'R_odds'/'B_odds.' Where R_ev/B_ev refers to the profit on a $100 credit winning bet and R_odds/B_odds refer to the odds that the fighter will win. Interestingly, plotting B_ev vs R_ev shows as the payout for a particular color increases it is more likely that fighters would lose or have an inverse relationship with fighter payout and winner. You can see this in figure 6. As R_ev or red payout increases, the winner is more inclined to blue and vice visa. Similarly, as the odds of a fighter decrease (toward negative) the more likely that fighter would win. In figure 8, looking at the bottom right corner, you see that as B_odds trend toward negative there are also more blue winners.

Lastly, even though a fighter's weight may be heavily regulated, physical differences are not uncommon. A fighter may have an advantage in reach, height, or age. Based on my knowledge, a longer-reach fighter will be able to punch the opponent better and manage the distance. You see this example through Jon Jones, who is regarded as the greatest UFC fighter alive and is one of the longest fighters in the light heavyweight division. A height advantage may not be so helpful because of the threat of wrestling takedown and jiu-jitsu submission. we would expect younger fighters to have more energy than older fighters, but the experience of an older fighter may be better. The resulting plot of blue physical attributes vs red physical attributes seems to confirm my assumptions.

Reach advantage seems to give a fighter a better chance at winning. In figure 8, B_reach vs R_reach, looking at the far-right side, we can see that R_Reach_cms generally wins over blue around 210 cm similarly the same observation can be made for B_reach_cms around the same reach. One thing to note, those observations are made at the very extreme. That means there must be a clear advantage in reach for a fighter to have an increased chance of winning.

It is hard to draw a conclusion from a height advantage. In figure 9, B_height vs R_height, it is hard to draw a conclusive trend with height advantage. Looking at the farthest right corner where the red fighter's advantage seems to indicate a winner but not so much the blue fighter.

Extreme differences in age seem to have given the younger fighter the advantage. In figure 10, look at the far-right side for younger blue fighters and the top side for the younger red fighter.

Methodology:

Data Cleaning:

From the initial inspection, there are three things that we must fix regarding the dataset: obsolete predictors, predictors of similar values, and missing values. To deal with each one of these, we will remove unnecessary predictors, perform some feature engineering by combining like predictors, and then fill in n/a values with the appropriate actions.

Firstly, we will remove predictors that won't contribute to the predictive model. The fighters' rankings will all be removed due to many missing values. Even though we could fill the missing values with 0, it won't be helping with the model due to an imbalanced distribution. Furthermore, some fight-specific predictors such as date, country, or location won't be helpful due to irrelevance in a 1v1 bout. See the list of removed predictors in table 3 in the appendix.

Secondly, we will combine similar predictors together. There are a lot of fight predictors that pertain to red and blue fighters that could be combined. For instance, 'R_wins' and 'B_wins' are combined into 'wins_diff,' the differences between the two fighters. This action is performed with all the fight and fighter's records. we will keep the combined predictor and drop the two related predictors. Between predictors removal and combining predictors, we have dropped the number of predictors from 119 to 50 while maintaining the essence of the dataset.

Lastly, we will fill all the missing values with 0. we made this decision because the new predictors now indicate differences in certain fight aspects and therefore zero would indicate no distinct advantage in either fighter. After these three processes, we now have the baseline dataset.

Model Training

Now that we have a clean data set, the next step is to train a baseline ML model and then compare it to four other models. In this case, we chose random forest as our baseline model because of its ability to utilize random decision trees which enhance accuracy and combat overfitting. However, the downside of random forest is to draw a concrete factor that influences the final prediction. The other four models are logistic regression, Linear Discriminant Analysis (LDA), ridge regression, and boost. The reason is as followed. Logistic regression, when used as a classifier, separates the binary classification of the winner by a 50% threshold which means it is easy to implement and less complex than other models. LDA is a dimensionality reduction technique that is good at distinguishing two groups which are helpful in our case. Ridge regression is like linear regression except for a regularization term that will keep coefficients as low as possible in our case we would like to see if the reduction in predictor would help the final prediction. Lastly, boosting model combines a set of weak learners into strong learners to minimize training errors which could give us a better predictive model overall.

Optimization

Models training is just one part of getting the best predictive result. Next, we will use the following optimization technics to attempt to get better accuracy results: standardization/normalization (data scaling), feature selections, and cross-validation.

Firstly, as we saw in our Data Set Sources, different predictors seem to have different scaling. We want to perform data scaling because we don't want priority to be given to a particular feature. Some ML models will give a better performance after scaling. However, ML models such as random forests and boosting don't require scaling because it doesn't use coefficients to determine the importance of predictors instead, they only care about the order of the predictors.

Then, we will attempt to use manual feature selection and a built-in automatic feature selection. For the manual feature selection, we will remove any predictor that is above 0.2 in correlation. After this process, there are 44 predictors left from the original 50. The automatic feature uses a sklearn SelectFromModel() library that selects the features based on importance weights. At the end of this process, we get 20 predictors down from 50 predictors.

Lastly, we will use 5-fold cross-validation which randomly separated our dataset into 5 sections (4 training sets and 1 test set). The purpose of cross-validation is to help us get better use of a limited dataset. In our case, CV won't really be necessary since we have 4896 observations which are adequate, but they should be tested.

Analysis Methods

The analysis method is simple. We will compare the accuracy score of 5 models with different optimization methods.

Analysis and Results:

Table 4: Summary of Models Accuracy Score

	Baseline Dataset	**Scaled Dataset**	**Manual Feature Selection**	**Auto Feature selection**	**Cross Validation**
Random Forest	64.5%	64.5%	61.1%	63.7%	62.8%
Logistic Regression	66.7%	66.6%	63%	66%	66%
LDA	66.1%	66.1%	63.5%	66%	66%
Ridge Regression	66.2%	66.4%	63.3%	66%	66%
Boosting	62.8%	62.8%	58.7%	63%	63%

Overall, logistic regression seems to perform the best at 66.7% accuracy using the baseline dataset. Boosting with manual feature selection seems to perform the worst at 58.7% accuracy. In general, most accuracy value maintains around 60%-67%. Moreover, manual feature selection seems to reduce all models' accuracy. Logistic regression, LDA, and Ridge regression seem to consistently outperform the other two models. The reason could be random forest and boosting average out from many models so the accuracy may not be high, but it is truer in the other three models. For the most part, the baseline dataset seems to perform better than the other optimization technics. Cross-validation seems to perform just as well as the auto-feature selection which is a surprise. For auto feature selection, we were expecting an overfitting model or a poorer fitting model when the number of predictors reduces from 50 to 20.

Conclusion/Finding:

Through our data exploration, we can conclude that a clear advantage in reach and age does seem to give an increase in winning. Height advantage is less conclusive with some red fighters seeming to show more wins with this advantage but not blue fighters. It seems like the saying 'the house always wins' holds true in our case. Betting odds and betting payout seems to be good indicator of the winner with a clear distinction between red and blue winners (Fig 6 & Fig 7). Lastly, it is possible for us to predict a fight with 66.7% accuracy using Logistic Regression using baseline non-scaled.

For future work, we could improve on data cleaning and feature engineering technics. The bulk of this project went into cleaning up the given data set and making sure it is good for ML training. Other options could have been used such as leaving red-fighter and blue-fighter data separate and replacing the missing values with average values or other methods. Additionally, we've only used 5 ML models in this paper. There are other classification models such as KNN or SVM that we could have tried. Overall, I think there are some things that can be done to increase the accuracy scores.

Lessons We Have Learned

- Start early: Working with real-world data is messy and you will quickly find your time stuck in data clean up.
- Be clear on what you want to find from your dataset: I quickly found myself exploring other aspects of the dataset and other correlations. This isn't a bad thing, but it just takes the focus away from the main goal.

Appendix:

Conor McGregor | UFC. 3 Aug. 2022, www.ufc.com/athlete/conor-mcgregor.

Get Started With XGBoost — Xgboost 2.0.0-dev Documentation. xgboost.readthedocs.io/en/latest/get_started.html.

History, Mma. "History of MMA." *MMA History - Mixed Martial Arts*, 16 Apr. 2019, mmahistory.org/history-of-mma.

History of UFC | UFC. 23 July 2021, www.ufc.com/history-ufc.

Jon Jones | UFC. 31 Oct. 2022, www.ufc.com/athlete/jon-jones.

Little, Becky. "How Did Humans Evolve?" *HISTORY*, 18 Jan. 2022, www.history.com/news/humans-evolution-neanderthals-denisovans.

"Sklearn.Discriminant_Analysis.LinearDiscriminantAnalysis." *Scikit-learn*, scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html.

"Sklearn.Ensemble.RandomForestClassifier." *Scikit-learn*, scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

"Sklearn.Feature_Selection.SelectFromModel." *Scikit-learn*, scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html.

"Sklearn.Linear_Model.LogisticRegression." *Scikit-learn*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

"Sklearn.Linear_Model.RidgeClassifier." *Scikit-learn*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifier.html.

UFC 270 | UFC. 18 Jan. 2022, www.ufc.com/event/ufc-270.

"Ultimate UFC Dataset." *Kaggle*, 9 Oct. 2021, www.kaggle.com/datasets/mdabbert/ultimate-ufc-dataset?select=ufc-master.csv.

** For the full list of all libraries used, please check the attached Jupiter notebook.

Table 2: List of Predictors in Dataset
['R_fighter','B_fighter', 'R_odds', 'B_odds', 'R_ev', 'B_ev', 'date', 'location', 'country', 'Winner', 'title_bout', 'weight_class', 'gender', 'no_of_rounds', 'B_current_lose_streak', 'B_current_win_streak', 'B_draw', 'B_avg_SIG_STR_landed', 'B_avg_SIG_STR_pct', 'B_avg_SUB_ATT', 'B_avg_TD_landed', 'B_avg_TD_pct', 'B_longest_win_streak', 'B_losses', 'B_total_rounds_fought', 'B_total_title_bouts', 'B_win_by_Decision_Majority', 'B_win_by_Decision_Split', 'B_win_by_Decision_Unanimous', 'B_win_by_KO/TKO','B_win_by_Submission', ‘B_win_by_TKO_Doctor_Stoppage', 'B_wins', 'B_Stance', 'B_Height_cms', 'B_Reach_cms', 'B_Weight_lbs',R_current_lose_streak','R_current_win_streak', 'R_draw', 'R_avg_SIG_STR_landed','R_avg_SIG_STR_pct', 'R_avg_SUB_ATT', 'R_avg_TD_landed', 'R_avg_TD_pct', 'R_longest_win_streak', 'R_losses', 'R_total_rounds_fought', 'R_total_title_bouts','R_win_by_Decision_Majority', 'R_win_by_Decision_Split', 'R_win_by_Decision_Unanimous','R_win_by_KO/TKO', 'R_win_by_Submission', 'R_win_by_TKO_Doctor_Stoppage', 'R_wins','R_Stance', 'R_Height_cms', 'R_Reach_cms', 'R_Weight_lbs', 'R_age', 'B_age', 'lose_streak_dif', 'win_streak_dif', 'longest_win_streak_dif', 'win_dif', 'loss_dif', 'total_round_dif', 'total_title_bout_dif', 'ko_dif', 'sub_dif', 'height_dif', 'reach_dif', 'age_dif', 'sig_str_dif', 'avg_sub_att_dif', 'avg_td_dif', 'empty_arena', 'constant_1', 'B_match_weightclass_rank',, 'R_match_weightclass_rank', "R_Women's Flyweight_rank", "R_Women's Featherweight_rank", "R_Women's Strawweight_rank", "R_Women's antamweight_rank", 'R_Heavyweight_rank', 'R_Light Heavyweight_rank', 'R_Middleweight_rank', 'R_Welterweight_rank', 'R_Lightweight_rank', 'R_Featherweight_rank', 'R_Bantamweight_rank', 'R_Flyweight_rank', ‘R_Pound-for-Pound_rank', "B_Women's Flyweight_rank", "B_Women's Featherweight_rank", "B_Women's Strawweight_rank", "B_Women's Bantamweight_rank", 'B_Heavyweight_rank', 'B_Light Heavyweight_rank', 'B_Middleweight_rank', 'B_Welterweight_rank', 'B_Lightweight_rank', 'B_Featherweight_rank', 'B_Bantamweight_rank', 'B_Flyweight_rank', 'B_Pound-for-Pound_rank', 'better_rank', 'finish', 'finish_details', 'finish_round', 'finish_round_time', 'total_fight_time_secs', 'r_dec_odds', 'b_dec_odds', 'r_sub_odds', 'b_sub_odds', 'r_ko_odds', 'b_ko_odds']

Tabe 3: List of Removed Predictor
'date','country','location','weight_class','gender','constant_1','finish','finish_details','finish_round', 'finish_round_time','total_fight_time_secs' + all of rankings

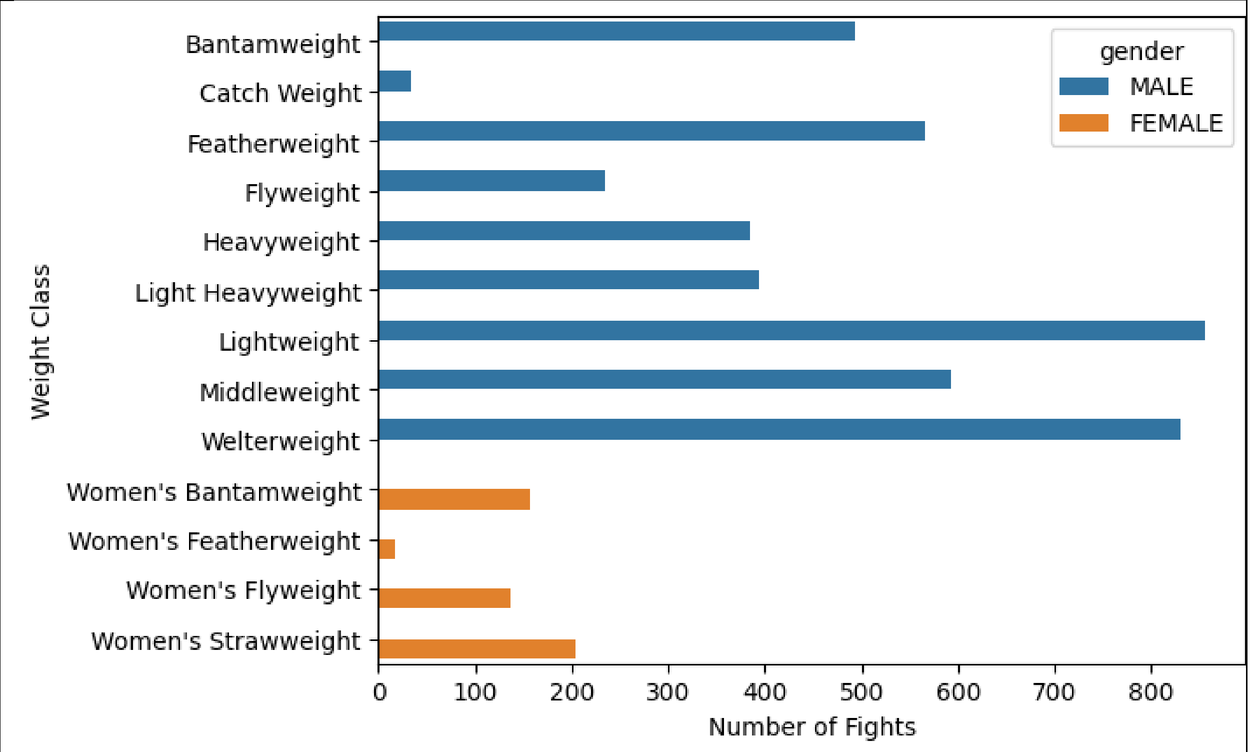


Figure 1: Number of Fight in each Weight Class

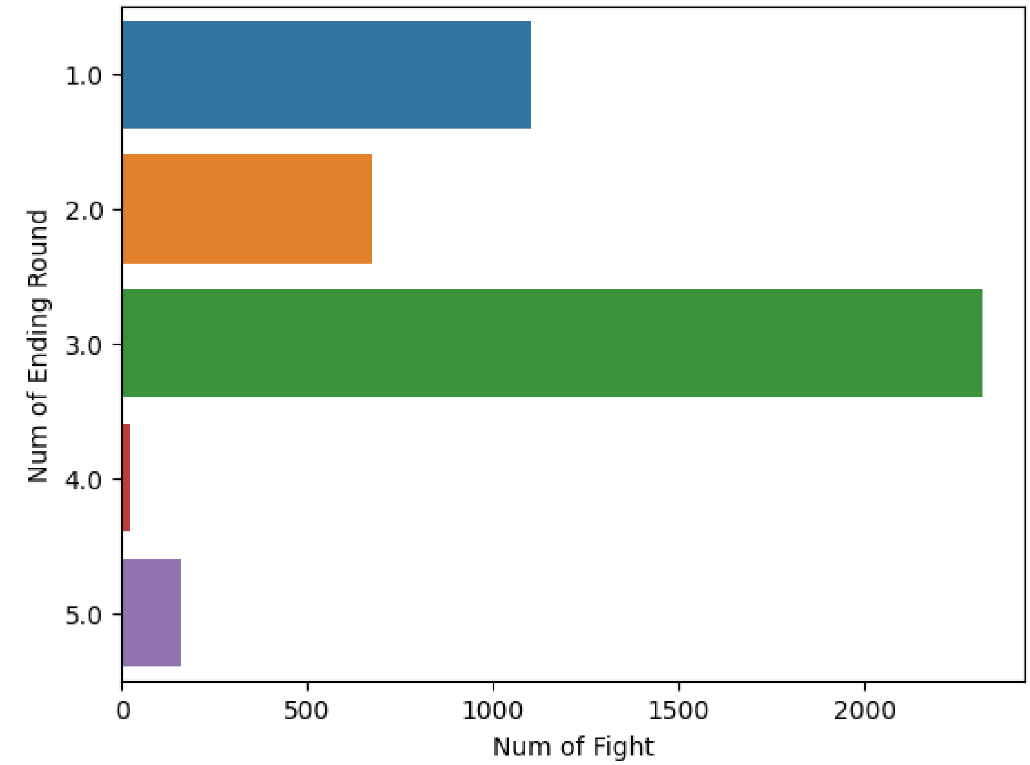


Figure 2: Number of Fight for each Ending Round

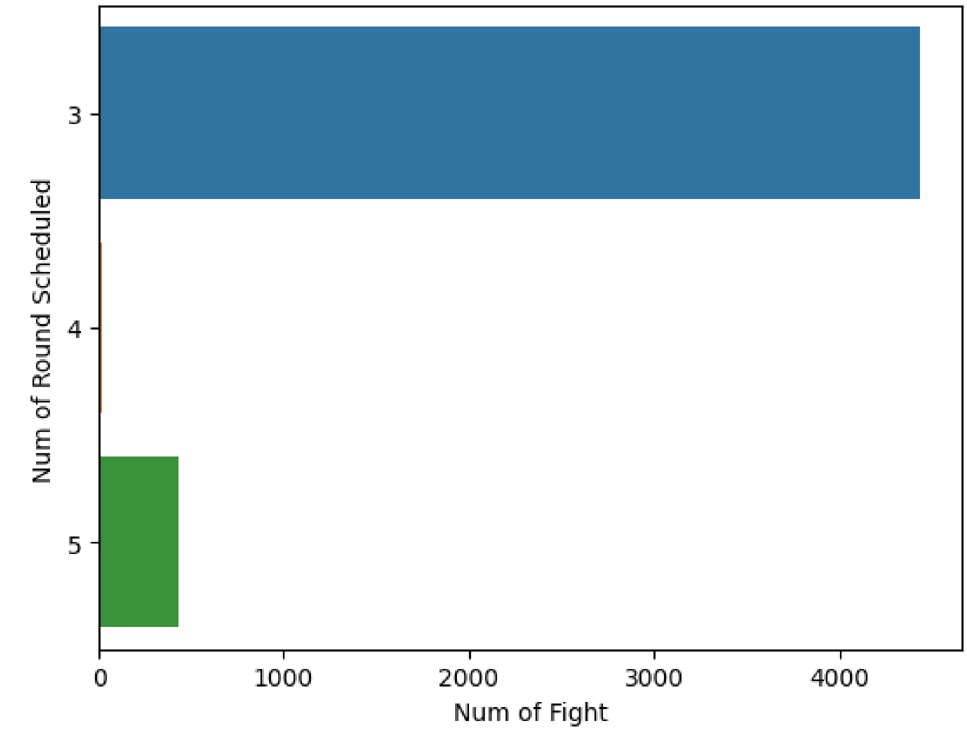


Figure 3: Number of Fight based on the number of Round Schedule

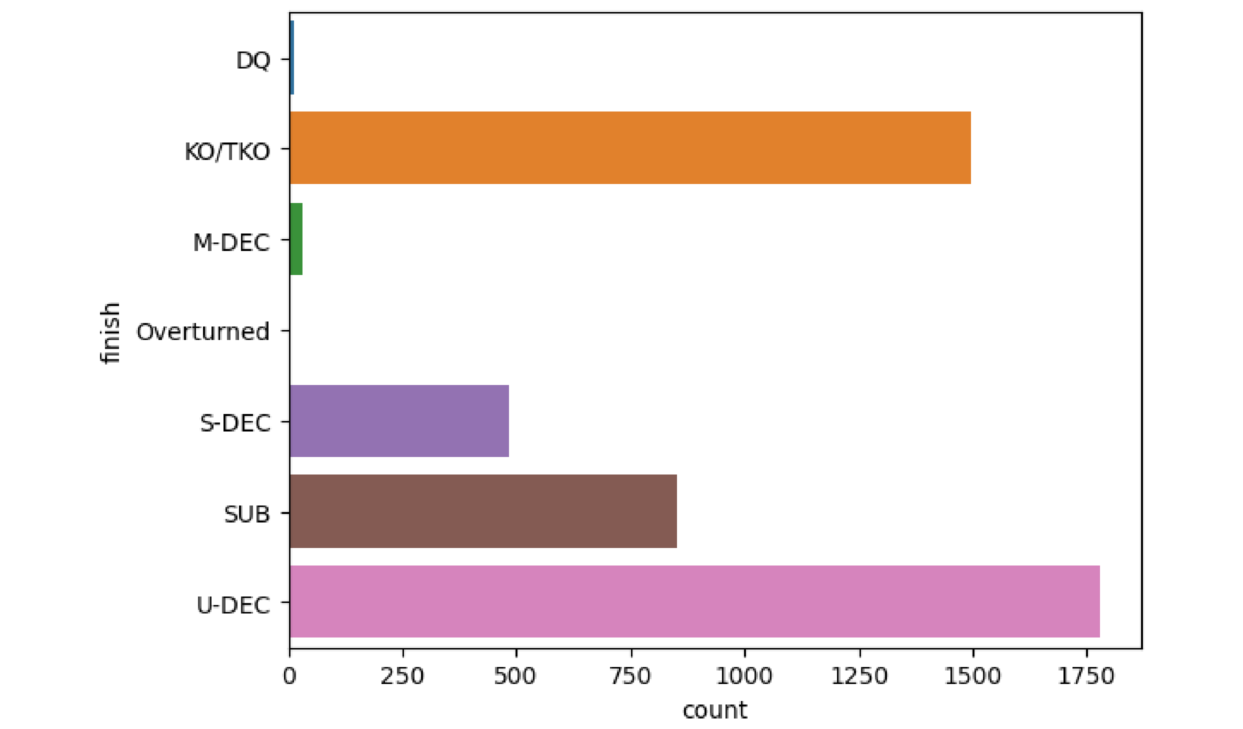


Figure 4: General Type of Finishes

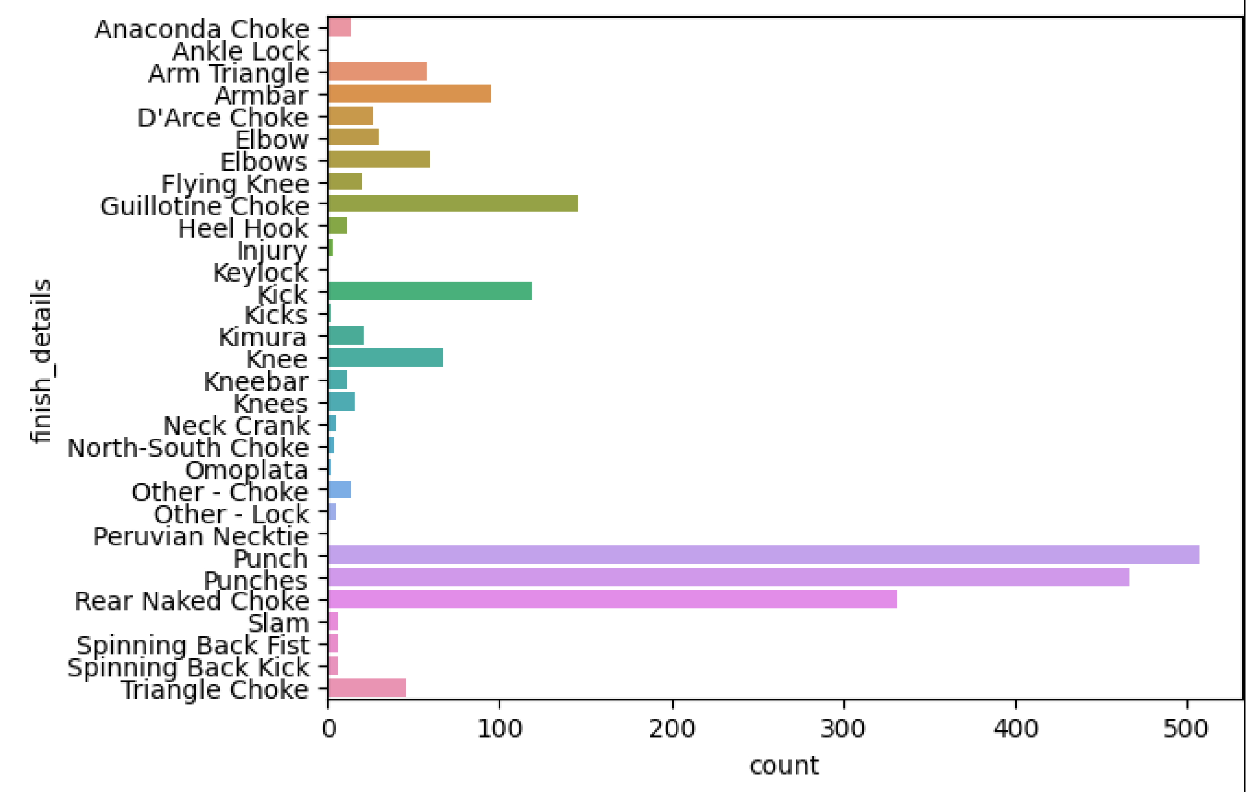


Figure 5: Specific Finishing Details

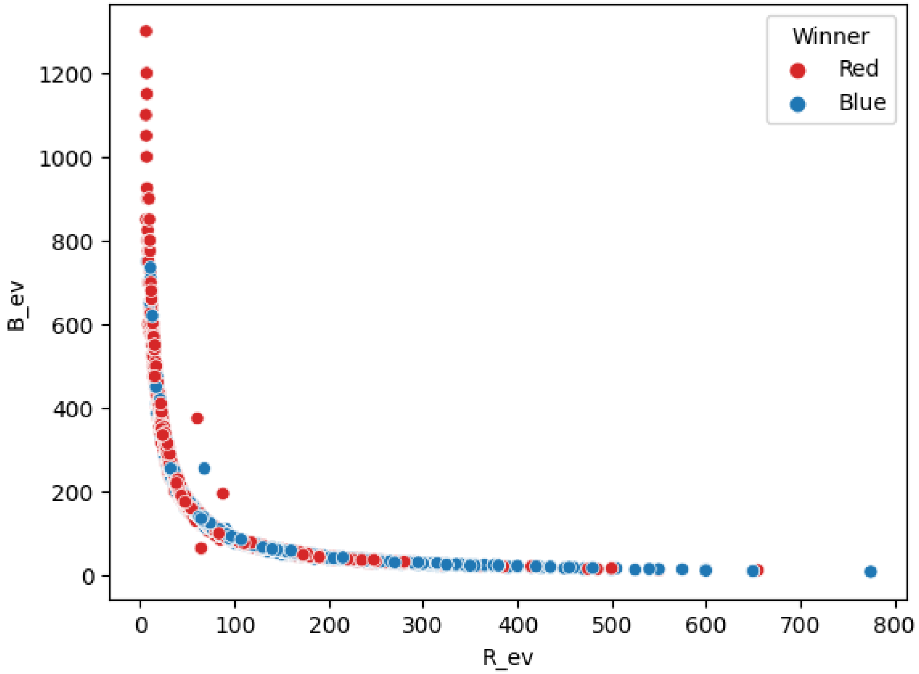


Figure 6: Betting return based on $100 bet

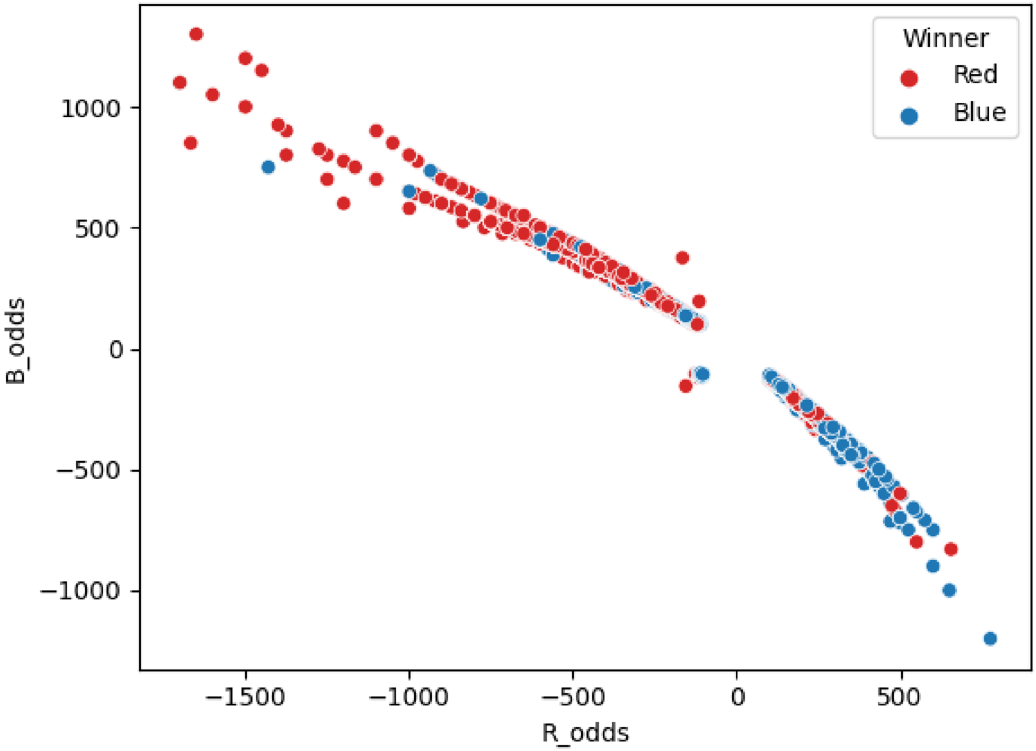


Figure 7: Betting Odds

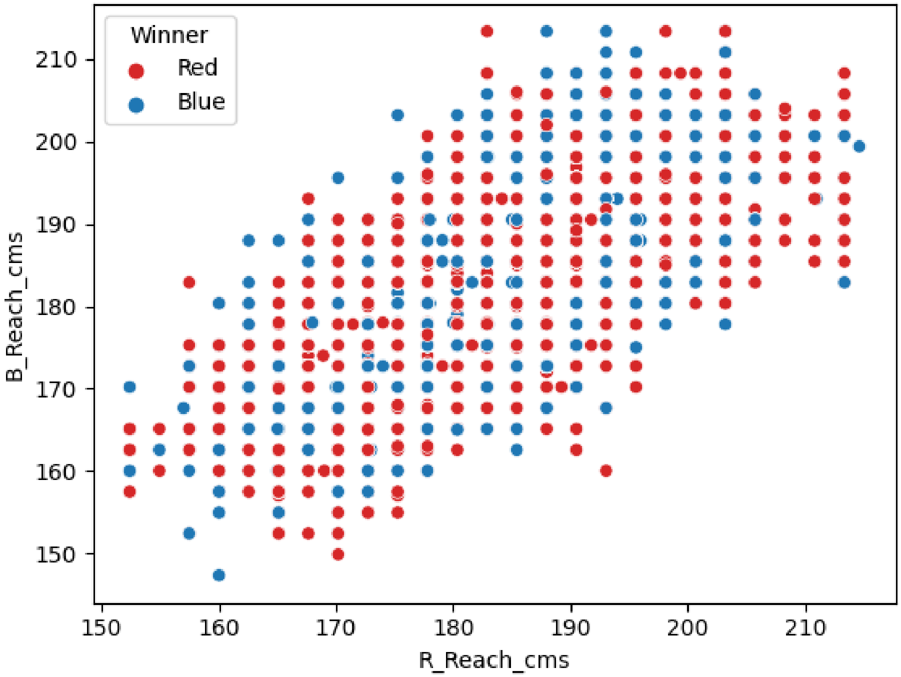


Figure 8: Blue Reach vs Red Reach

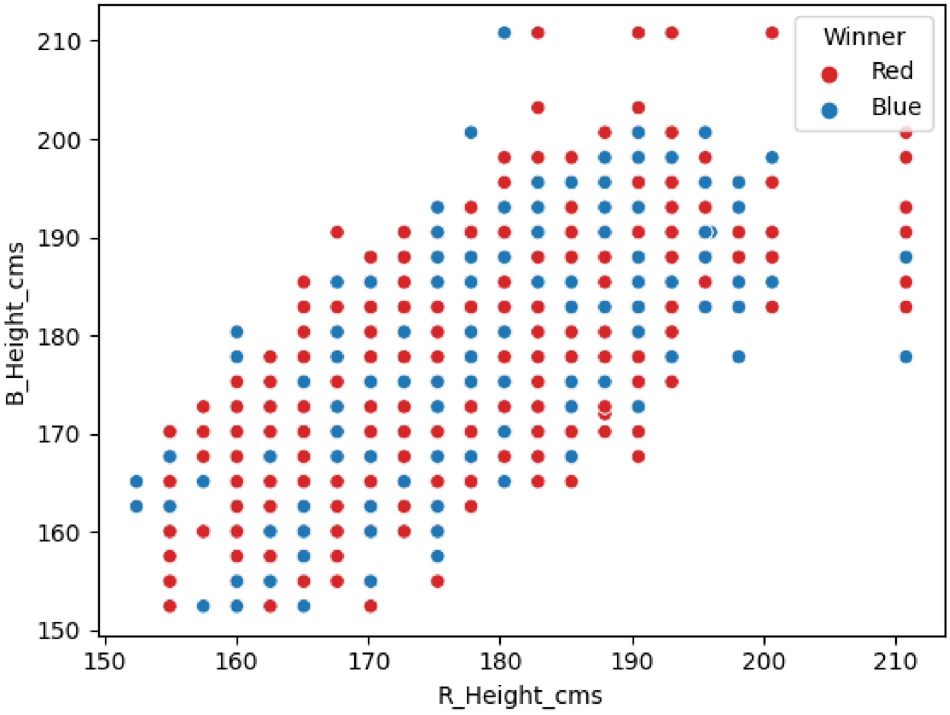


Figure 9: Blue Height vs Red Height

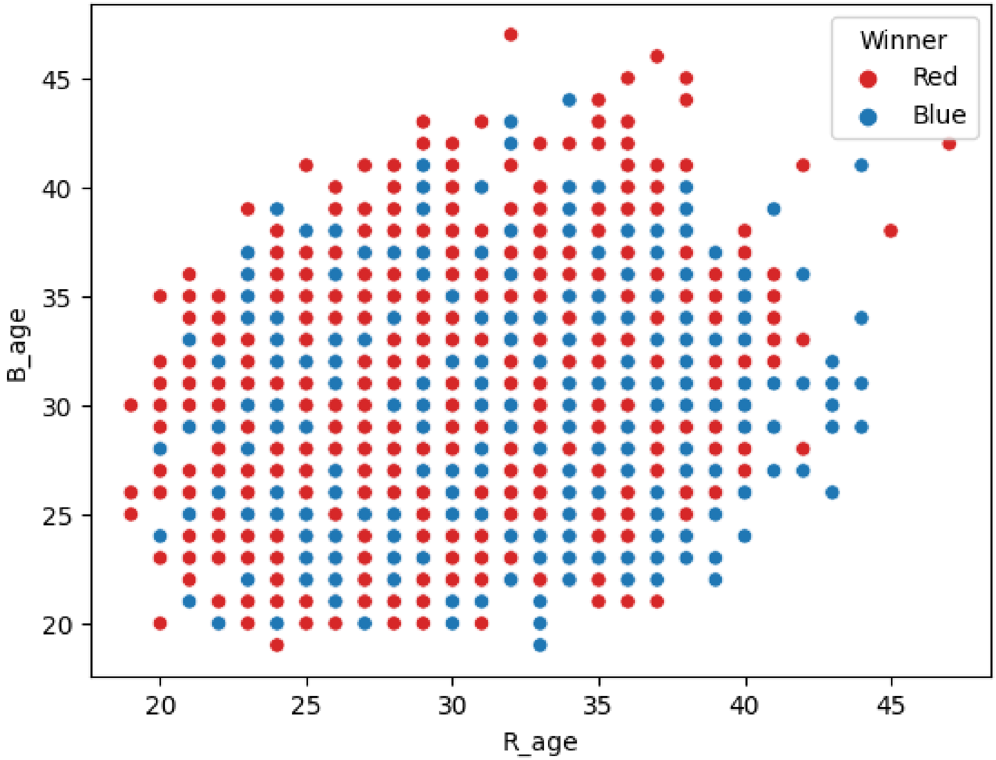


Figure 10: Blue Age vs Red Age