A case study on UFC: What makes a UFC Champion?

Introduction

The Ultimate Fighting Championship (UFC) is the largest Mixed Martial Arts promotion company in the world. Established in 1993, it has grown into a $7bn enterprise as of 2018[1] and has popularised MMA along the way into the mainstream sport that is known today. The organisation features the top tier of MMA athletes and attracts fighters globally with the substantial pay-out of money, sponsorships and fame that is attached to the UFC Champion title. UFC's MMA incorporates all legal combat styles and techniques governed by the Unified Rules of Mixed Martial Arts. It is highly volatile and unpredictable and as popularity for MMA and UFC increases, conducting an analysis can be worthwhile in producing insights to the technicality of the sport and explore the path to becoming a UFC champion. This can be useful to UFC and sponsoring partners as they can observe specific areas of the sport and focus on strategies aimed to further promote their brand image.

Analysis domain, questions, and plan (500 words):

Data Source

The project will utilise data on the UFC sourced from the Kaggle repository which were originally scraped from the UFC stats website. This will be a dataset of UFC fight records dating from 1993 – when it first established – until the most recent point in 2019 where the dataset was updated. There are 5144 rows where each row represents one fight event and 145 features. Within those features are metrics that records both fighter's past performance in fights; such as offensive strikes landed, submissions attempted, fight record etc., as well as fighter's characteristics; height, weight, stance etc., and also fight details; date, location, winner, whether it's a title bout and so forth. Given the features, we can identify this dataset as a machine learning binary classification problem, where prediction of which fighter will come out on top can be performed. We aim to do something similar in this study but also delve into understanding the sport of MMA and the UFC using this dataset.

Questions

With this study, we aim to address some important problems and provide clarifications to some questions in yielding a better understanding of the sport of MMA within the UFC association. Typically, analytical problems pertaining to sports in general are identified as two categories: on-field and off-field analytics[2]. On-field analytics is essentially performance

---

[1] MMAjunkie 2018, Dana White says UFC now 'worth $7 billion' after ESPN deal, USA Today, viewed November 25th 2019. https://eu.usatoday.com/story/sports/mma/2018/08/20/dana-white-ufc-brand-worth-7-billion-espn-deal/111249212/

[2] The Evolution and Future of Analytics in Sport, 2017, Proem sports, viewed 25th November 2019. https://www.proemsports.com/single-post/2017/06/22/The-Evolution-and-Future-of-Analytics-in-Sport

analysis[3] – understanding how athletes can improve performance through a data-driven approach. On the other hand, off-field analytics is concerned with the business side of the sport. UFC and business sponsors would be highly interested in this as this would be their considering factors for their marketing strategies.

These problems will serve as the foundation for our study and the scope of this study will be bound by addressing the identified questions below:

- How has MMA within the UFC evolved in the past 26 years of existence?
- What are the advantages of an athlete's biometrics in influencing performance, and thus, the outcome of an event/fight?
- Can we yield the highest performance in a fight?

Upon addressing these questions, we will have explored the evolution of MMA within UFC and the study should provide us bountiful amount of insight that can help to answer what makes a UFC champion.

Plan

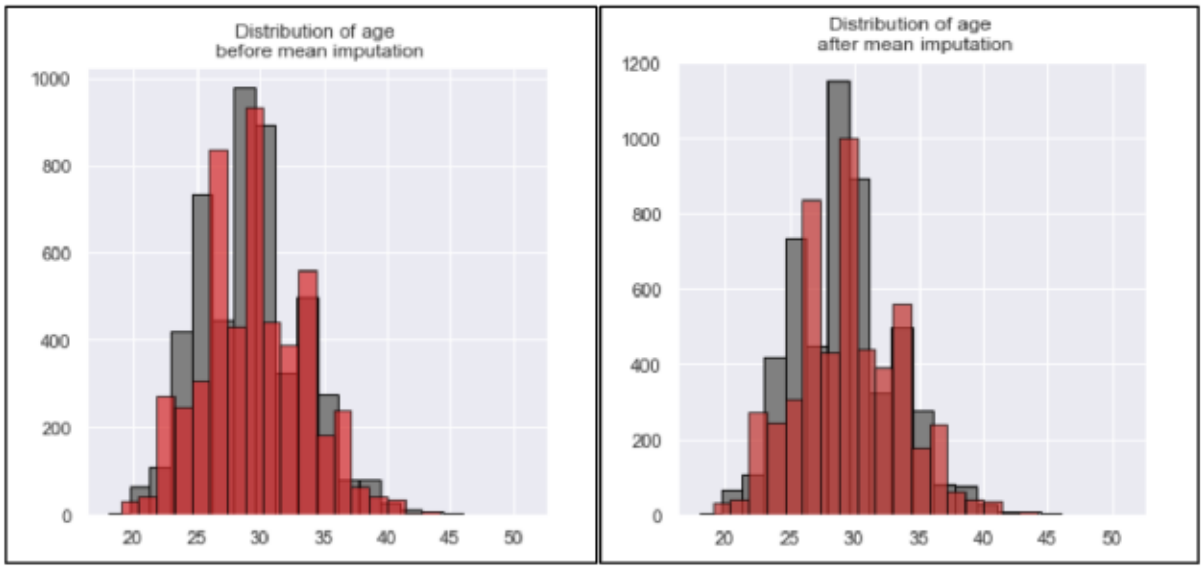
The following approach will act as guide in this study:

1. Obtain the relevant data needed for the study; concatenate and/or remove variables that are invalid to our study
2. Perform data cleansing in preparation of our study; address null values and detect outliers
3. Initially investigate into data to observe any characteristics
4. Feature engineering (if necessary)
5. Perform exploratory analysis on data
6. Attempt to model the ultimate UFC fighter, looking at the

Findings and reflections (1000 words):

Since there were some initial data wrangling steps to perform, decisions had to be made in the best way to impute missing data. Specifically, we looked into the distribution of age as we will analyse the feature further in the study. A normal distribution for age was identified so we had the decision to use the mean or the median to impute null values. There was hardly a difference in the two and so the decision was to use the mean. After imputation with the mean, the distribution was unaffected.

[3] Performance Analysis, Sport Northern Ireland, viewed 25th November 2019. http://www.sportni.net/performance/sports-institute-northern-ireland/performance-science/performance-analysis/



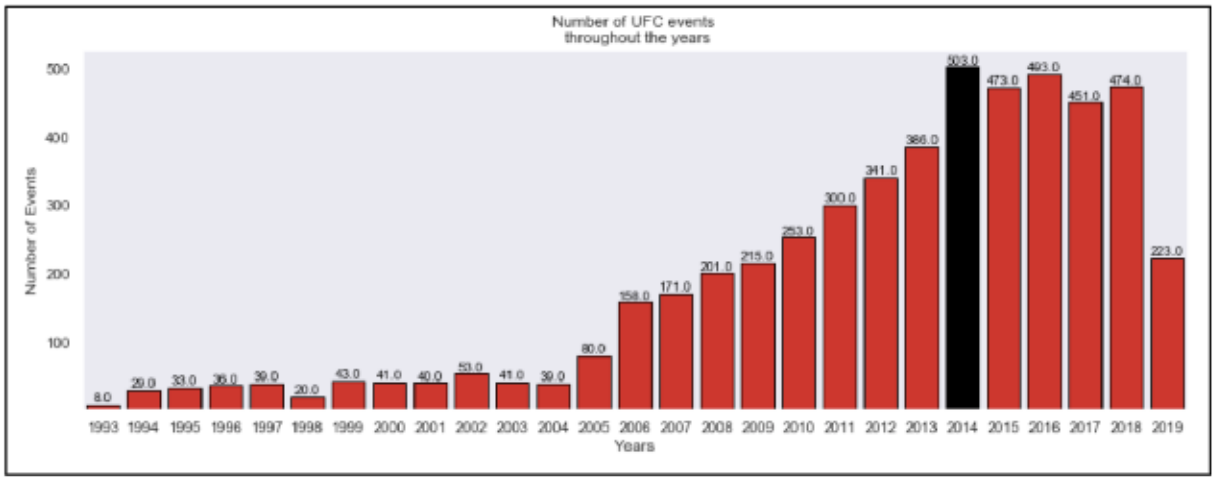
We begin our study investigating in the evolution of the UFC. Deconstructing the structure of the UFC, many progressive developments from the beginning of its time are observed that are beneficial to current and potential athletes and from a business standpoint.

The inclusivity of female athletes in the UFC has grown. From its beginning as an all-male dominated association, now 6% of all athletes have been women representing the UFC. The progression was first observed in November 2012 when Ronda Rousey became the first female fighter to sign. The first female weight division was then initiated a year later providing the opportunity for female fighters to become UFC champions. There is not a better time for female athletes to compete in the UFC than now. This is also an opportunity for female-oriented businesses to use this platform to elevate.

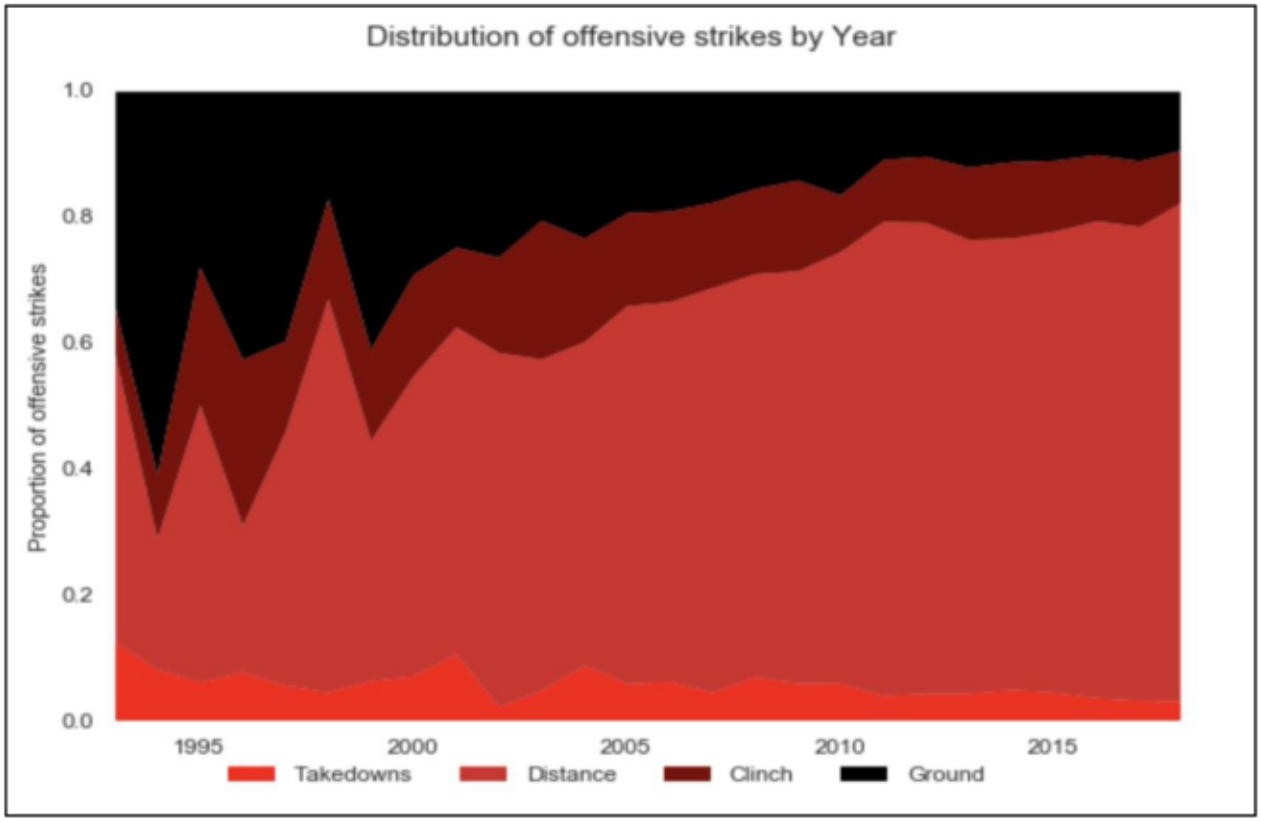
We then investigated the weight divisions and found by far the most competitive divisions were lightweight and welterweight with many fighters having fought in the two divisions before – since most fighters are liable to switch between divisions by making weight. As more competition equals high risk which means high rewards, athletes tend to gravitate towards these two divisions for that reason. We also observe a defunct open weight which no longer serves a purpose in the modern UFC. This was when fighters of different measurements would compete with one another and so there was a strong handicap to the underdog.

By country, we see that the US overshadows other countries in the popularity of UFC. This comes as no surprise as the UFC is based in Las Vegas, Nevada, US. But we do observe upcoming secondary markets where it would be advantageous for UFC to continue tapping into developing their brand such as in Brazil, Canada, the UK, Australia and Japan.

Overall, from its humble beginnings, we see that the UFC has indeed continue to expand the popularity of MMA globally. Number of organised fights have consistently increased with a peak in 2014. Expectations for the UFC to continue will certainly help any brand looking to expand their reach and also in enriching the competitive environment for athletes in becoming a top-tier athlete.



Looking at the fighting environment, we investigated the change in the offensive fighting styles over time. Modern fighters tend to perform more distance strikes than in the past. We see a contraction of the ground game where about 40% of all offenses were on the ground in the beginning. Clinch and takedowns meanwhile have stayed fairly consistent. Fighters need to understand how the modern game works and this is important for them to fit into the modern game.

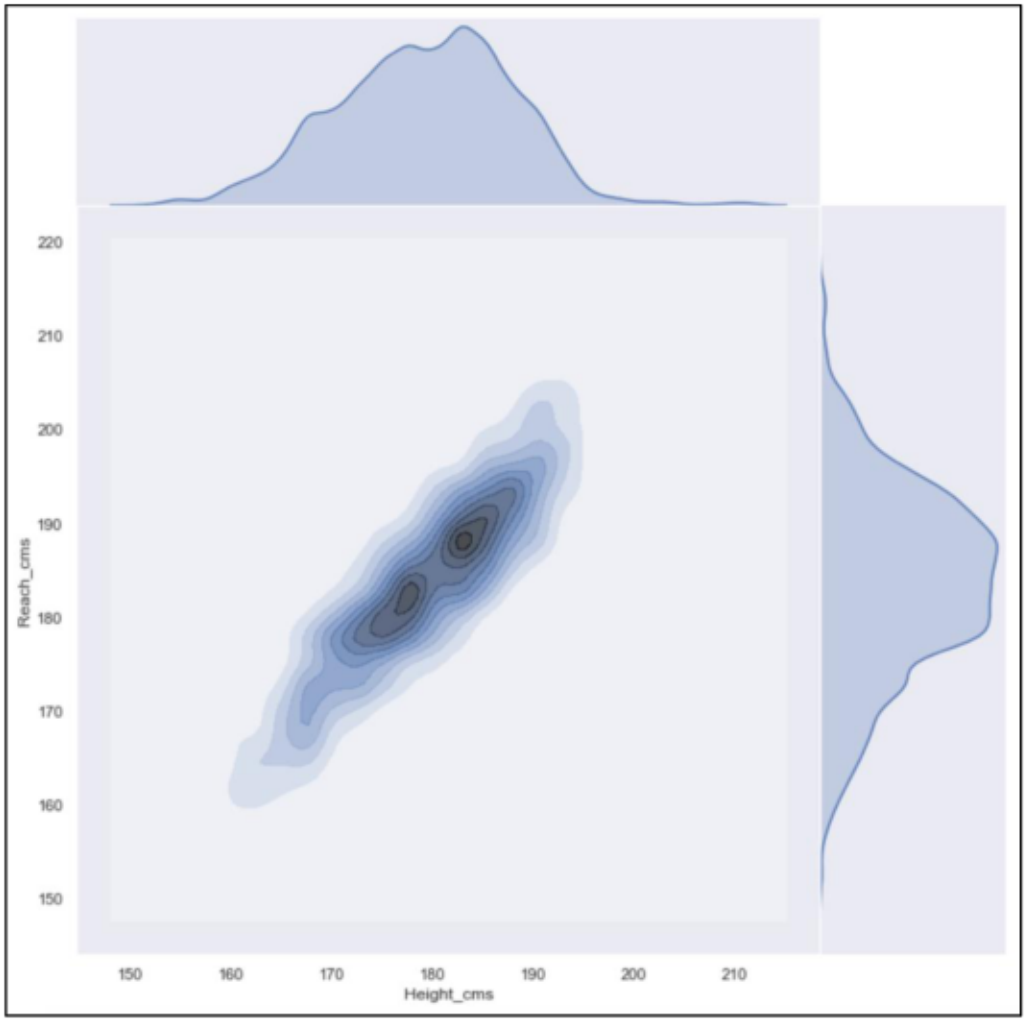


We also observed the breakdown of wins which provided us general information that the most common way for fighters to win is to KO/TKO an opponent. Moreover, there is no absolute change in the winning styles of athletes over time. Divisional wins also show nothing interesting in particular.

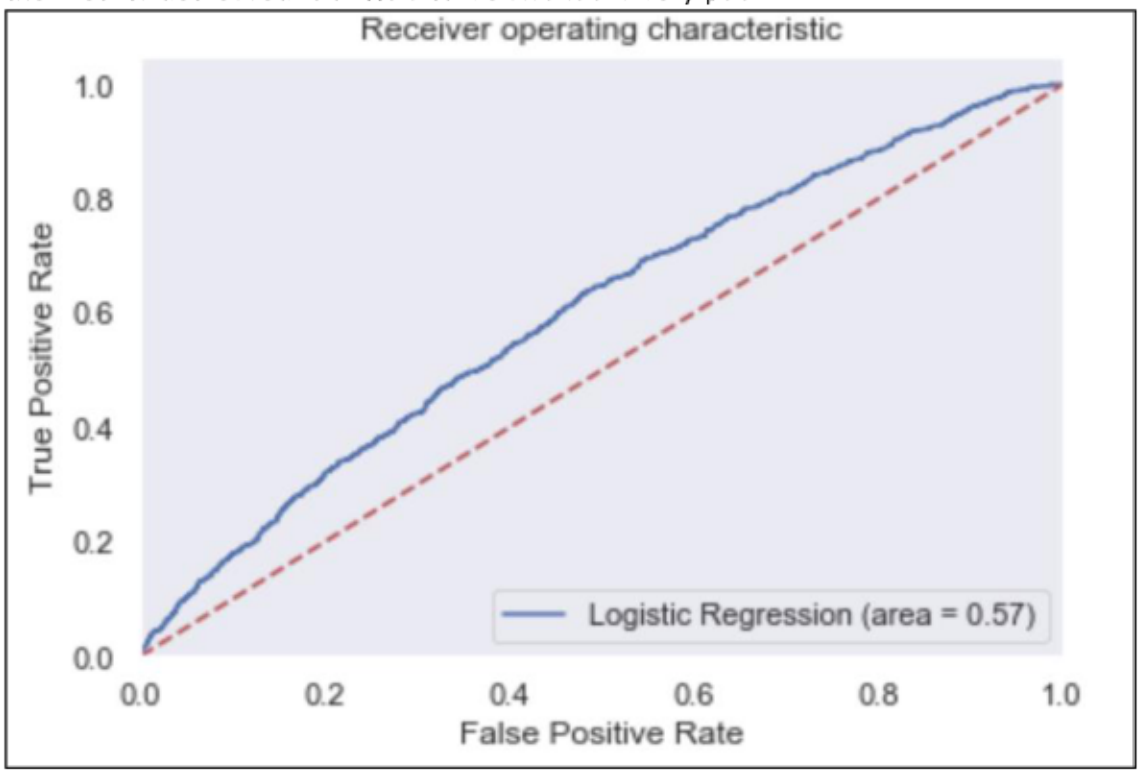
The study will now focus on uncovering an athlete's metrics that would give them an advantage over others.

Since age is an important factor in performance in any sport, we will first look at age. Our findings show that most UFC fighters peak around their late 20s to early 30s with a mean age of 29.3. This is consistent with most sports where athlete's peak performance is observed between the ages 26-35. We delve further to find out 79% of athletes are under the age of 35 in the UFC. Since UFC is an extremely physical sport, most fighters past their prime and the age of 35 are more likely to sustain serious injuries if they continue to compete.

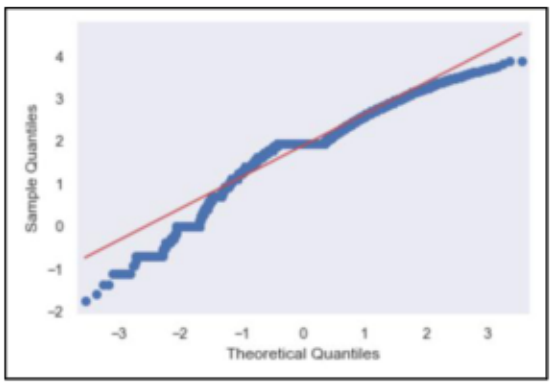
An athlete's height is also advantageous in a fight. Our findings show that height and reach share a linear relationship – that is, the taller you are the longer reach you have. We plotted a KDE plot to display the majority of the fighter's measurements in our data and find that most are clustered within the middle (darker sections). When we compared the mean difference of fighters on one side to the fighters on the other side, we see that there is a positive height reach difference when they win and negative when it's a loss.

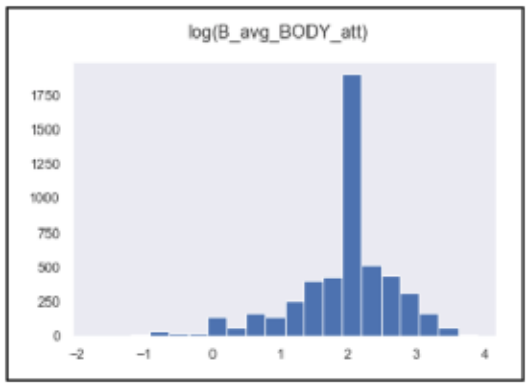


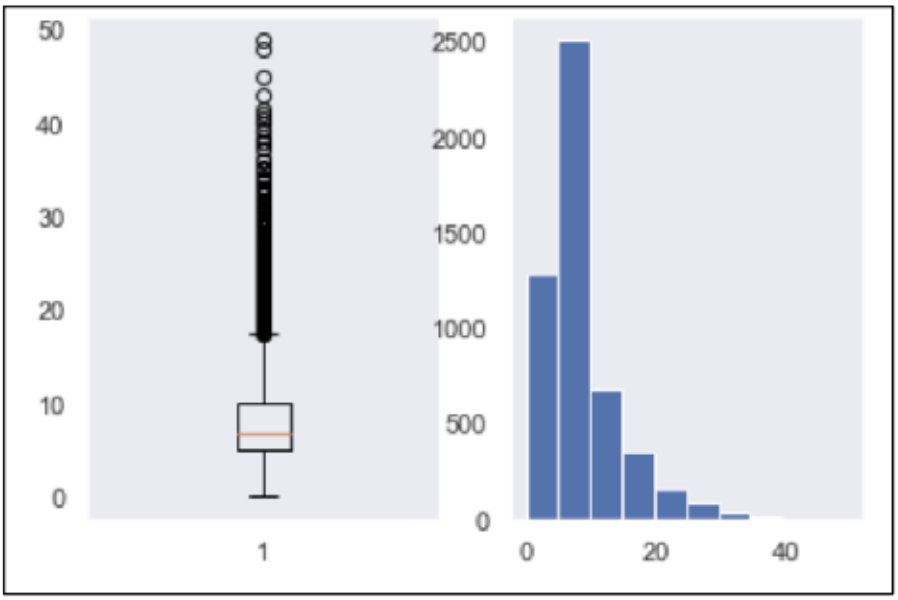
Our study now takes us to modelling the fight outcome with the data. For this binary, classification task, we will be using a logistic regression. Looking at pre-modelling statistics, the outcome is class balanced. Once we've derived some dummy variables and cleaned up dataset, we begin our modelling and evaluate using the confusion matrix. We find that the accuracy and f1 score is low (58%). We try to improve this by analysing a correlation matrix for correlation and multicollinearity in features. Removing the correlated features, we find the metrics decreased. Our ROC curve also looks very poor.



For the scope of our study, we will not be optimising the model but many things can be done to improve on this analysis. Fundamentally, the nature of our dataset can be transformed as looking at the Q-Q plots for the fight features, we can see that it is heavily left skewed. Log transforming it seems to improve a little but there are too many missing values in the dataset (~20%) and imputing seems to generate heavy outliers (as seen from the boxplot) in the dataset which ultimately affects our analysis.






50
40
30
20
10
0
1
2500
2000
1500
1000
500
0
0
20
40