Look Into UFC Fighter Data

KAI KHASU, Marquette University

This paper goes through my process of how I completed my first Data Science predictiono' analysis. A reader will learn a little about the history of statistics in sports and then how some statistics and graphs can be applied to a basic data set.

ACM Reference Format:
Kai Khasu. 2019. Look Into UFC Fighter Data. 1, 1 (December 2019), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Everyone will agree in saying that technology is improving at an astounding rate. So fast that people who are looking towards the future are saying that in 30 years, not knowing how to code will be equivalent to not knowing how to read today! Similarly, Data Science is a specific field that is booming just as fast!

For years now people have been using data science methods to predict the outcome of sporting events and improve their teams. The most mainstream, pop-culture reference to this would probably be the Brad Pitt movie, Money Ball. Everyone is trying to find the most predictive model that uses all the right features to predict an outcome of an event. Machine learning is making advances and people are hoping on that train.

Before we can get there however, we have to start simple. I started my search for The Model by learning the data analysis process. Throughout this paper I will be addressing each step throughout the process and all the different things I learned along the way. This process has not been easy and I ran into road blocks almost immediately. UFC is a relatively young industry. It was founded in 1993 which only gives it about 26 years of history. While there have been plenty of fights to record data since then, there have not been as many reputable publications to pull valuable information from. While martial arts has been around since what feels like the dawn of time with karate and samurais and whatnot, unfortunately I never came across an easily readable paper within the scope of this project.

Instead I turned to other fields with the same idea in mind. I found important history dealing with baseball, basketball, football and soccer. I tried to translate what I learned from each back to the UFC but it was not an easy feat. UFC is not like most sports. As far as I know, it is not a sport many people grow up dreaming of making it big in. It is not for everyone and takes a very disciplined person to master. This will become more clear throughout the paper.

Going into this adventure of cleaning data, I went in with a few things in mind. First I tried to gain exposure to the different areas of the pandas library. Second, I wanted to see if there was any correlation between the number of punches a fighter threw and the number of wins they had. I was looking to see if there was a sweet spot of the ideal amount of punches. Finally, the last thing the I was looking for was if there was an ideal body type or physical attributes that winning fighters shared.

Author's address: Kai Khasu, khasukai@gmail.com, Marquette University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Association for Computing Machinery.
Manuscript submitted to ACM

Before getting into the body of the paper, I will leave you with food for thought throughout the paper. UFC is not like most sports. It isn't the same as basketball as in you have to put a ball in a basket and another team prevents you from doing so. Its not like baseball where you try and hit a ball thrown by a pitcher out of the park. It is more like chess. When an fighter does something the other fighter must react appropriately all of the time. It's cliche, but true, one wrong move can literally be the end of a fighters career.

2 LITERARY REVIEW

In preparation for and throughout this project, I spent a lot of time reading up on Bill James and his idea of Sabermetrics. James defined Sabermetrics as "the search for objective knowledge about baseball" through statistics and other empirical baseball knowledge. In 1982 Bill James introduced the world to his idea of Sabermetrics, and revolutionized how sports organizations were run. [2] Since then, the application of some type of Sabermetrics can be found in almost any sports environments! I mean that all sports analysts are trying to find some 'truth' behind the sport.

This search for truth is not easy. There is no industry standard as to what is the correct way to get to the truth, however statistics have been the path of choice of many people practicing Sabermetrics. With the advancements in tech and availability of computers, anyone is able to pick up a computer, learn some basic R and/or Python and give it a go. However, this is where the true art comes into play.

Sabermetrics is not the practice of collecting all the data that is available and feeding it to a computer. Rather it is discerning what information is most important and valuable. Truth is not defined as the culmination of everything present. Rather to get to the truth of something, you have to ask the right questions about the matter at hand.[2]

Although Bill James introduced Sabermetrics for baseball the theory has been applied to all different sports. One of the most common places to find people 'trying to find truth in a sport' is the Casino floors in Vegas. People will grow up playing sports their entire life and think that they know something that nobody else knows about for a sport. Take basketball for example. Say someone realized that by wearing a headband, less sweat gets in players eyes and they will make more shots. They can then go to their computer search for the teams with the most players that wear headbands and then place all their money them! It's that easy.[4]

The problem arises when that person loses all their money... People in Vegas have dedicated their lives to finding these things called 'edges' which are their own theories about what is true about a sport. Generally these theories are a little more predictive of an outcome than a headband.

Along with reading Bill James's Sabermetrics ideas, parts of Dean Oliver's Basketball on Paper: Rules and Tools for Performance. While I did not read much of this book, I learned more about 'asking the right question'. In Oliver's book he declares that there are 'Four Factors' which have a huge impact on the game. These factors are shooting percentage, turnovers per possession, offensive rebounding percentage, and getting to the foul line. Notice none of these have to do with wearing a headband!

Essentially, what Oliver is saying here is that these are a baseline for where to start your basketball model. That baseline is then meant to be build upon and made more complex and advanced to improve the prediction model. The goal of a professional sports better is to achieve an accuracy score of 52.4%.[3] If this can be achieved, at the typical Las Vegas spread odds of -110 or 11/10, the better will always at least break even. There is some pretty simple math that proves this it is easily seen on https://www.thesportsgeek.com/sports-betting/math/.

Billy Walters is another interesting character in the gambling world for having a model that is the gold standard for sports betting. He has never had a losing month in the last 30 years which is unheard of in the gambling world. I don't necessarily agree with this argument but throughout my work for this project I have come across people trying to use

variations of the Anthropic Principle to prove that some perfect model does exist. While I wont say that, I will say that because people like Billy have had such success, that there is a way to predict to an extent the outcome of an event but there is no perfect model that predicts everything.

3 METHODS

When it came time to clean and organize my data, I chose to go with python and pandas to clean it. The data consisted of a lot of information that I was not really sure what to do with so I split this into two sets.

3.1 Learning Set

This section will explain the initial dirty work I went through to get familiar with pandas. Initially, like anyone knew to a language, I had no idea what to do. I did really know what my data was telling me and with the little understanding I had, I had no idea what to do with. I wanted to look at what my data was telling me so I did simple mean and sum functions on my data frame. With the information i got with those simple functions I made some initial graphs like Fig. 1. Although this doesn't tell much it was a step in the right direction.

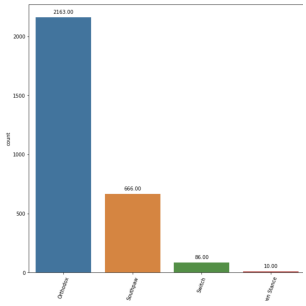


Fig. 1. An countplot of the stances fighters use

3.2 Applying What I Learned

This time around, with a better understanding of what data I had, I was able to create much more informative graphs like Fig. 2. Figure 2 is a scatter plot of the percentage of significant strikes a fighter landed and colored by whether or not they won. I repeated this scatter plot with the different pairs of fighter statistics. I compared each fighters attempted and landed head shots and body shots as well as significant strikes like fig. 2. Many of the graphs turned out to look very similar to the one in figure 2. I will go more into why I think this was the way that it was in the results.

4 RESULTS

4.1 Before Regression

Before I applied any regressions, I made my biggest discoveries when I was initially exploring the data. I found Figure 1 for example which showed the almost all of the fighters fight following the open stance method. I also found that almost all fights end in the third round in decision, unless they are title fights. Those fights go 5 rounds and then end in decision.

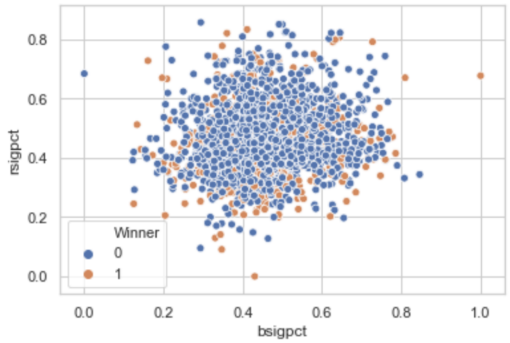


Fig. 2. A scatter plot of each figher percentage of signifiant strikeslanded

Like I said in the methods section, many of my significant findings were a lot like figure 2. I was unable to discern anything from the graphs. This was how it was for the relationships between blue fighters (x) percentage and the red fighters (x) percentage. x would be head strikes or body strikes.

For a while I was stuck at this point. Before I applied any statistic regressions to see what was statistically significant, all I had were a ton of graphs that looked like figure 2. I decided that although there was no relationship, that does not mean the data isn't telling you anything.

4.2 Interesting Relationships

Here are a few relationships I found that may not really provide much beneficial information but they are still important because they show information about how the fights tend to turn out.

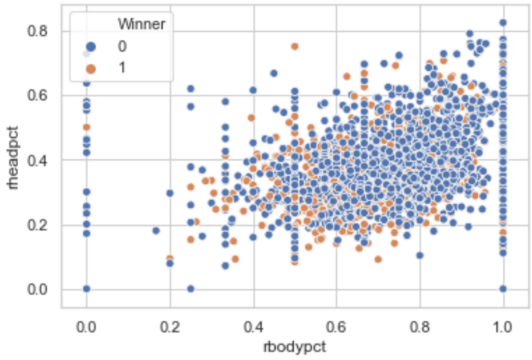


Fig. 3

Like I said, Figure 3 doesn't really show much except for the linear lines that show up throughout the graph. Most of them fall on 'normal numbers'.

Figure 4 shows the distributions of fighters ages based on if they won or lost. This is interesting because although it doesn't prove that a fighter of a certain age will win, it shows that fighters who win tend to be about ages 26 - 32

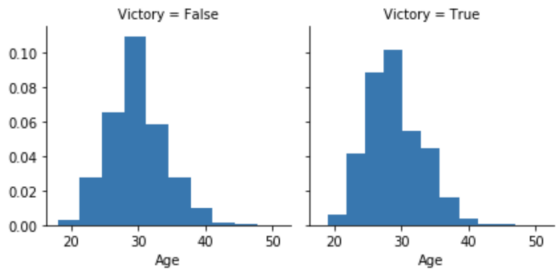


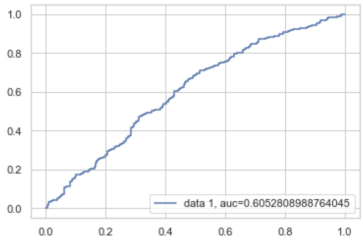
Fig. 4

4.3 After Regression

Because I wasn't really able to make any inferences off the scatter charts or even the regression plots, I decided to do a little research and apply some more advanced statistic based measurements on my data. I started with a simple logistic regression using stastmodels.api Logit function trying to predict the winning fighter. I did this a few times until all the variables I ran the regression with were significant or had a p-value less than 0.05. These regressions revealed that the most significant variables in my data set were the losing fighter's perctage of head stikes that hit, the winning fighters previous opponents average body strike percentage, the winning fighters previous opponents average head strike percentage and the losing fighters previous opponents average head strike percentage.

Following this trend of the metrics of my data, I decided to see how my regression would do in a Confusion Matrix. It did poorly... the resulting Matrix was [[494, 6], [258, 9]]. If that doesn't make sense, the accuracy of my program classifier was 0.65, the precision was 0.6, and had a recall value of 0.03. It performed abysmally.

Finally, in a last attempt to try something new, I included an AUC ROC curve to top things off. The purpose of this is to 'indicate how well the probabilities from the positive classes are separated from the negative classes'.[1]



here I included the AUC ROC graph for you to see how bad it is. it recieved a value of 0.6 which is 0.1 more than what the website I learned about this from called "worthless".

5 DISCUSSION

Now we ask the question, "What does any of this mean for my 'truth' in UFC?". And really it means we have found nothing. We know that most people fight Orthodox fighting style, and almost all fights go to decision.

Although our AUC ROC showed our predictions are really really bad, all of this was not for nothing. Just because I wasn't able to predict any significant factors in the data that was available to me does not mean there is nothing to find.

It just means I was not looking in the right places. That is one of the conclusions I came to before I started applying logistic regressions to my info-less scatter plots. I realized that just because there is no relationship between the axes on the scatter plot doesn't mean the points on the plot don't mean anything.

I decided that there was such a concentration of points in certain areas because that's how fighters were most comfortable fighting. Most of my data was for Orthodox Stance fighters so it would make sense that most of the data supported a common trend of everything averaging about 9 head shots a round and 9 body shots a round. A fight is one-on-one. There is no second person on your team helping you do what you do. It's not like basketball where if a point guard is playing poorly the center can carry the team. In the analogy I'm trying to make Orthodox is the point guard and well call South Paw a center. And what I am saying is, just because Orthodox fighting isn't working doesn't mean a fighter can just switch to South Paw.

I think this failure to predict is a great indicator that it takes more than numbers to predict how a fight will go. A lot of people will say that decisions should be made without emotion or a 'gut' feeling. While this may be true, there is some level of understanding of a sport that goes into finding what is important. For example, when Dean Oliver wrote his book about basketball, he understood basketball because he was a part of the legendary Dean Smith's coaching team at North Carolina. He saw what worked on a daily basis and knew what to look for when he was doing his analysis.

Moving forward into future projects, I will look to know more about my data than just the number and letters that make it up. If at all possible, I will try to experiment in the real world with the data I am analyzing. This attempt really taught me that when the columns of a pandas data frame are just numbers, they're really kinda useless. It isn't until those numbers have some meaning or weight to them that they can really be manipulated and understood.

At the beginning of this process, one of my goals was trying to predict if the physical attributes of a fighter, like height and age, had impact on the result of the fight. I wasn't able to find any variables that proved to be statistically significant. I tried to see if the average number of strikes a fighter throws and has had thrown against him impact the outcome of a fight. And in the end, although I had variables that proved to be significant, they weren't very good at proving anything.

In the future, if I were to continue with this project, I think that I would learn more about what is important when fighting. Like in basketball I know turnovers, rebounds and points are all important. When it comes to UFC, I have no experience in the octagon. I can read that take downs are important and not getting hit is important but I can't tell you how important they are when it comes to the flow of a fight. So I think with a little more free time to dedicate to this project, I could have made a lot more discoveries about the data I working with.

6 CONCLUSION

Now looking back on my decision to try and predict the outcome of a UFC fight, I realize that I jumped into the deep of a pool end without the proper flotation devices. While I learned a lot throughout the process, I wish I had chosen something simpler. Like trying to predict the prices of a stock as it fluctuates throughout the day. No, that was a joke. I do wish I had done something I was more familiar with, like predicting the score of a basketball game. I guess this was also a good lesson in that not all experiments end with the result you expect. I'm not saying I was expecting a model that was 100% accurate, however I would have liked a model that would have told me something. So that will be a project for this Winter Break.

REFERENCES

[1] [n.d.]. *Let's learn about AUC ROC Curve!* Retrieved December 8, 2019 from https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152

[2] Phil Birnbaum. [n.d.]. *A Guide to Sabermetric Research | Society for American Baseball Research.* Retrieved December 10, 2019 from https://sabr.org/sabermetrics/single-page

[3] The Sports Geek. [n.d.]. *Sports Betting Math - How To Win Money at Sports Betting.* Retrieved December 8, 2019 from https://www.thesportsgeek.com/sports-betting/math/

[4] Dean Oliver. 2011. *Basketball on Paper: Rules and Tools for Performance Analysis* (2nd. ed.). Potomac Inc.