UFC Fight Predictor and Analysis of the Relationship Between MMA Statistics and Winning Fights

Louis Romeo

University of Arizona

CSC 496 Principles and Practices of Sports Analytics

David Lowenthal

12/9/2024

Abstract: The UFC is the largest and most premier MMA organization in the world, holding events in more than 30 countries and broadcasts spanning over 40 languages. With this rise in popularity the outcomes of world title and ranked divisional fights has become increasingly important to the state of the sport. On the ufcstats.com website there are 8 tracked statistical categories across all fighters having completed at least one bout in the UFC. Through the use of both linear and non-linear methods, I created a Python project which contains programs that attempt to answer the overall question, can UFC statistics help predict a fight winner?

Introduction: Using the UFC's tracked statistical information, I first sought answers to the simpler hypothesis, can statistics tell us anything about the UFC? I believed that finding or not finding any correlations between a certain fighter's attributes would be the rightful starting point to get any predictive insight. Along with some implied findings (more strikes attempted results in more landed strikes per minute), I was able to find the strongest correlation between striking statistics and fighter success (higher ranking number) rather than wrestling, meaning that generally better strikers are favored in the rankings. Moreover, when looking at fights that go the full distance, it is likely that a fighter with better striking statistics (mainly higher number of significant strikes) is favored by most judges to win the decision. Using this knowledge, I created multiple programs with the purpose of predicting a hypothetical fight between any two fighters in the same division, collecting data from the top 15 in all 7 men's divisions for analysis. I started with a linear predictive algorithm, calculating a winner solely on statistical advantages. As I progressed I created a metric called "performance score", which factors in all 8 statistics, gives each statistic a weight, giving striking stats slightly more weight than wrestling for example, and outputs a winning percentage when given input is two fighters in the same division. Through

tweaking the performance score metric and the given weights, I compared certain fight predictions against the betting line percentages for fights that have already taken place for accuracy. Lastly, I experimented with some non-linear models, including random forest and gradient boosted modeling to observe more into winning and losing correlations.

Background: The use of statistics in combat sports is prevalent but can be considered minimal compared to other major (mostly team) sports. Still, combat sports statistics and metrics can provide important data on a fighter's striking/wrestling capabilities and more. With data on ranking numbers, physical attributes (age, height, reach, etc.) and statistical categories such as strikes and takedowns landed and attempted per minute, we can compare fighters and see potential advantages and disadvantages in any MMA category. Currently the most predictive metric on MMA fights is the betting lines which are developed by oddsmakers. Given that those numbers can be converted to percentages, they can provide insight on who is more likely to win before a fight.

Implementation: Using a variety of Python programs, I created a user-friendly interface that when run with the properly cleaned data asks the user to create either a scatter plot, boxplot or run regression statistics between any two tracked metrics to look for correlations. As for the linear predictor, I simply tracked all statistics and modified the user-interface to select a division and the names of any two fighters to see their statistical advantages. When using the winning margin I used my created performance score metric and converted the difference between the scores to a percentage and continued to modify the weights for accurate percentages. The models I used for nonlinear prediction utilized random forest and gradient boosting, and was also able to

output a ranking of all fighters in the division by their percentage chance of winning against anyone in the division (first is 100%, last is 0%). In a separate file, I also implemented a data visualization builder that outputs a heatmap, graph of other divisional statistics, and more.

Results: When searching for correlations between statistics I found in large part that there was not any definitive correlation between most statistics. However I did notice that the amount of strikes absorbed per minute decreases drastically as the divisions increase weight (likely because of strike power), the number of attempted takedowns also decreases as divisions increase and more trends, most of which can be interpreted in the context of sport. With my fight predictor, I was able to produce console output with direct statistical comparisons given any two names and comparing advantages.

```
Enter the folder path containing division CSV files (e.g., ufc_stats): ufc_stats
Enter the division (available: Bantamweight, Featherweight, Flyweight, Heavyweight, Lightweight, Light_heavyweight, Middleweight, Welterweight): Middleweight
Enter the first fighter's name: Khamzat Chimaev
Enter the second fighter's name: Sean Strickland

Statistical Comparison:
Metric              Khamzat Chimaev     Sean Strickland     Advantage
--------------------------------------------------------------------------------
Rank                3                   1                   Sean Strickland
Age                 30                  33                  Khamzat Chimaev
Reach (in)          75.0                76.0                Sean Strickland
SLpM                7.89                5.45                Khamzat Chimaev
Str. Acc.           60%                 41%                 Khamzat Chimaev
SApM                3.98                3.9                 Khamzat Chimaev
Str. Def.           55%                 63%                 Sean Strickland
TD Avg.             3.0                 1.0                 Khamzat Chimaev
TD Acc.             50%                 60%                 Sean Strickland
TD Def.             85%                 85%
Sub. Avg.           1.5                 0.2                 Khamzat Chimaev
Last Fight Result   W(SUB)              L(DEC)              Khamzat Chimaev

Comparison Results:
Khamzat Chimaev Advantages: 7
Sean Strickland Advantages: 4
Advantage Categories for Khamzat Chimaev: Age, SLpM, Str. Acc., SApM, TD Avg., Sub. Avg., Last Fight Result
Advantage Categories for Sean Strickland: Rank, Reach (in), Str. Def., TD Acc.
Predicted Winner: Khamzat Chimaev
```

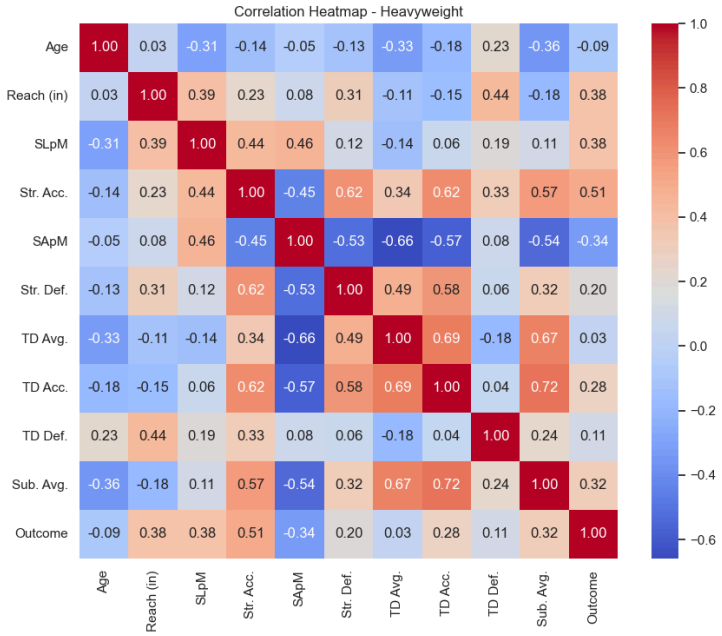
When predicting a victory margin the program will output the divisional rankings sorted by winning chance against any given opponent or “the field”, which I found to be a very interesting ranking system given that the fighter at the top has the advantage in most categories.

Bantamweight Division:

Division Analysis (Sorted by Winning Chance):

	Rank	Fighter Name	Winning Chance (%)
2	2	Umar Nurmagomedov	100.000000
1	1	Sean O'Malley	97.026847
0	C	Merab Dvalishvili	81.665554
10	10	Ricky Simon	78.407098
3	3	Cory Sandhagen	67.415441
11	11	Adrian Yanez	64.172001
14	14	Jonathan Martinez	59.757318
4	4	Marlon Vera	59.186713
7	7	Rob Font	57.985439
9	9	Song Yadong	54.847111
12	12	Raoni Barcelos	41.137570
13	13	Chris Gutierrez	40.371758
8	8	Dominick Cruz	37.263461
5	5	Deiveson Figueiredo	34.860913
15	15	Douglas Silva de Andrade	15.880783
6	6	Pedro Munhoz	1.000000

When generating charts for divisional analysis the most interesting chart I found to track possible correlations was a heatmap, that shows a slightly increased correlation between some statistics depending on the division (figure shows heavyweights have a higher takedown accuracy and submission average correlation than most other divisions).



Conclusion: Overall the goal of my project was to take a deeper look at UFC/MMA statistics and track any trends or patterns for the purpose of predicting a fight. With the rapidly increasing popularity of combat sports and competition using statistical analysis can be an important aspect for prediction. Through the use of plots, charts, linear and non-linear prediction, and statistical analysis I was able to develop a predictive insight into UFC matchups. As for the future of the project, I will continue to refine the performance score, use betting data to confirm model accuracy and continue to track possible trends in the statistics to find any more fight winning metrics. Through my passion for MMA and my knowledge of sports analytics I will keep working hard to utilize statistics in the UFC in the same way as other major sports.