2024

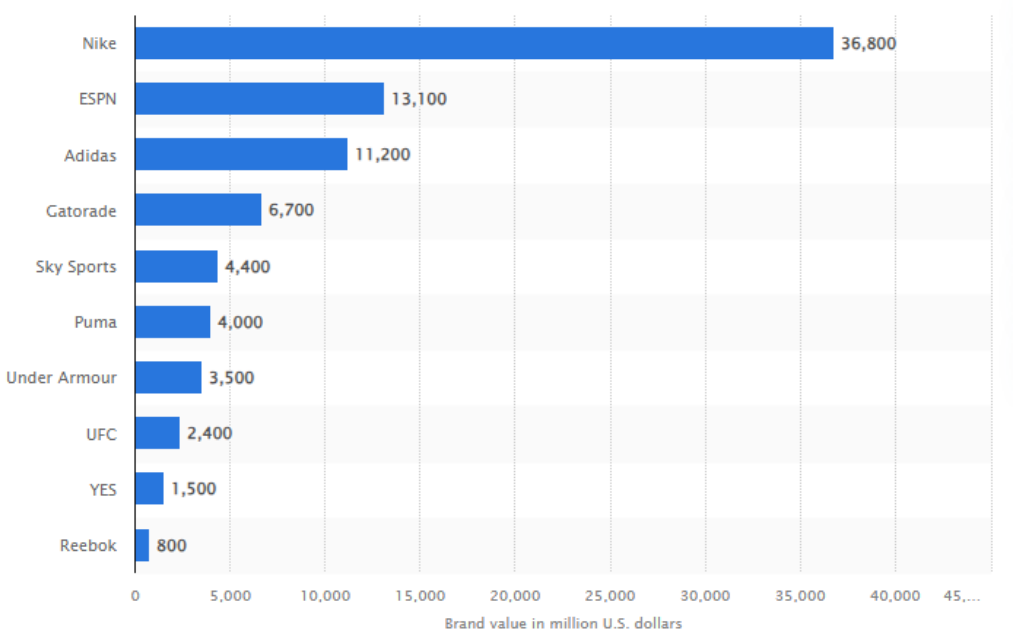
UFC Fight Prediction Group Project Report

TEAM 007
ALEJANDRO FORERO ENG
JULIETTE NOELLE WONG
SAMUEL WADE WANG
ANDREW BERKOWITZ
GITHUB

Contents

Background

The Ultimate Fighting Championship (UFC) is the most well-known mixed martial arts organization in the world today. ting in 1993, it has grown its way to $600+ million revenue from 2013 - 2016, having a single year growth of 335% ("Topic: Ultimate Fighting Championship (UFC)."), and being one of the top 10 most valuable sports business companies in 2019 ("World's Most Valuable Sports Business Brands 2019."). As of August 2022, their highest live attendance was greater than 57,000 ("Most Attended UFC Events 2022.") and are averaging 447,000 pay-per-view buys in 2018 ("Topic: Ultimate Fighting Championship (UFC)."), with their highest buys being 2.4 million ("Topic: Ultimate Fighting Championship (UFC)."). With the UFC's current broadcast deal being with ESPN, as of December 2023, mixed martial arts was the #9 sport in consumer interest ("Interest in Sport Types in the U.S. 2023."). It is clear to see that the UFC played a major role in bringing mixed martial arts to the forefront of the sports world.



This figure shows 2019 top 10 most valuable sports business companies ("World's Most Valuable Sports Business Brands 2019.").

With the popularity of mixed martial arts and a league-dominating viewership, it is only natural that sports betting is following. With the myriad of fight statistics, it would benefit both individuals and sports betting companies to better understand what factors drive specific outcomes and how to better predict those outcomes. This helps the individual place and win their bets and companies to better estimate the odds. After understanding individual statistics on what drives outcomes, we can pivot to explore if those statistics and fighters have an impact on the number of pay-per-view buys. Having a sense of how to sell more pay-per-views would ensure that the UFC company continues to grow.

Literature

Our team reviewed three pieces of literature associated with our topic. In the first two papers, the authors ran multiple types of machine learning algorithms to predict fight outcomes. Even though the process of both authors were the same, they came to different conclusions. Drwismer's 2,258 sample dataset had a stochastic gradient descent regression normalized model that had a prediction accuracy of 86.4% ("Predicting UFC Fight Scoring - Multivariate Linear Regression."), while McKinley McQuaide's 3,355 sample dataset had a 61.23% accuracy. McQuaide noted that the red fighter wins 62.6% of the time, which would result in better accuracy than the model created "Applying Machine Learning Algorithms to Predict UFC Fight Outcomes."). The disparity between both authors' results is interesting and could be from multiple factors - different fight data, varied factors, or even errors in the models. Having a baseline comparison from these two papers lets our team know that the data is random, and we should not be surprised if we get a high or low accuracy. The third paper we reviewed used past fight and betting odds data to create an optimal betting strategy using an artificial neural network. While this paper was less relevant to our problem, it inspired us to create a table showing the records of a fighter leading up to a match.

Team Process

The team is following a traditional and straightforward process. We have two separate datasets that need to be cleaned and transformed into four datasets. After the data is clean, we will have the ability to join them together if necessary and begin running machine learning algorithms. This will give us initial insights into how factors interact with each other and influence an outcome. We plan to incorporate models such as decision trees, random forest, logistic regression, and linear regression with a train/test split or cross-validation to avoid overfitting. We will evaluate the individual models and identify the most appropriate to answer our questions of "Who will win the fight?" and "How to sell more pay-per-views?".

Progress

The team has worked together to research the topic to fill in any knowledge gaps, learn about what others have achieved, and plan out a process for how to approach the problems we are trying to solve with the data available to us. Currently, we have cleaned out data into usable formats to answer our questions. The method we used to clean the dataset was to create separate datasets that would be easily joined together if needed. The datatypes of many columns in the main dataset, ufc_data_till_UFC_292.csv, were character, so during our process we had to convert them into usable datatypes. Our secondary dataset, ufc_ppv_buys.csv, needed some cleaning to fix inaccurate data and data that was incompatible with the primary dataset.

R_fighter	B_fighter	R_KD	B_KD	R_SIG_STR.	B_SIG_STR.	R_SIG_STR_pct	B_SIG_STR_pct	R_TOTAL_STR.	B_TOTAL_STR.	R_TD	B_TD
"character"	"character"	"integer"	"integer"	"character"	"character"	"character"	"character"	"character"	"character"	"character"	"character"
R_TD_pct	B_TD_pct	R_SUB_ATT	B_SUB_ATT	R_REV	B_REV	R_CTRL	B_CTRL	R_HEAD	B_HEAD	R_BODY	B_BODY
"character"	"character"	"integer"	"integer"	"integer"	"integer"	"character"	"character"	"character"	"character"	"character"	"character"
R_LEG	B_LEG	R_DISTANCE	B_DISTANCE	R_CLINCH	B_CLINCH	R_GROUND	B_GROUND	win_by	last_round	last_round_time	Format
Referee	date	location	Fight_type	Winner							
"character"	"character"	"character"	"character"	"character"							

Year	Month	Day	UFC_Number	Opponent1	Opponent2	PPV
"integer"	"integer"	"integer"	"integer"	"character"	"character"	"integer"

The top figure shows datatypes of our main dataset and bottom figure shows datatypes of our secondary dataset.

The first dataset we created was fighter records, based on our main dataset. Our main dataset only had a column with the winner of the match. To get the match's loser, we used an if-statement equivalent to "if the winner column is the red fighter, then the loser is the blue fighter, or else the loser is the red fighter". This would give us the losers' names that would be stored in a vector, converted into a data frame, and then joined back into the main dataset. With winner and loser, we were then able to create a new dataset of the fighter, winner, loser, and date of the fight. Since we wanted to use a fighter's winning or losing streak as a potential predictor of PPV sales or fight outcome, we grouped by each fighter and sorted the rows by date to calculate the cumulative wins, losses, and draws before a match on a given date. We were able to use the streak_run function to get a winning or losing streak. This dataset allows us to look at a fighter's wins, losses, draws, winning percentage, winning streak, and losing streak ahead of a given match.

date	fighter	total_wins	total_losses	total_draws	win_percentage	win_streak	lose_streak
<date>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2023-08-19	Pedro Munhoz	10	7	2	0.588	1	0
2023-08-19	Amanda Lemos	7	2	0	0.778	2	0
2023-08-19	Mario Bautista	6	2	0	0.75	4	0
2023-08-19	Denis Tiuliulin	1	2	0	0.333	0	1
2023-08-19	Austin Hubbard	3	4	0	0.429	0	1
2023-08-19	Gerald Meerschaert	10	8	0	0.556	0	1

This figure shows the finished fighter record dataset.

The next step was to clean up our main dataset consisting of individual fight stats. When inspecting the dataset, most statistics were in string format. To deal with this, we used the separate function on the columns to split the first and second numbers and convert them into integers.

R_SIG_STR.	B_SIG_STR.
17 of 35	25 of 35
30 of 53	16 of 55
95 of 156	66 of 130
10 of 15	5 of 6
23 of 55	31 of 59
29 of 42	22 of 50

R_SIG_STR_LND	R_SIG_STR_ATT
17	35
30	53
95	156
10	15
23	55
29	42

The left figure shows example columns having statistics in a string. The right figure is the result after separating.

There were some characters in percentage columns that we had to deal with. We noticed the "---" characters seemed to symbolize 0%, so we replaced those when applicable. The next step for this dataset was to convert columns into usable datatypes. This included changing percentage columns from character to numeric, where we divided by 100 to get the decimal form. It also included changing time columns from minutes and seconds in character type to total seconds numerically. Since our main dataset had categorical columns, we created indicator variables for columns we wanted to keep using the dummy_cols function in the fastDummies package.

Once we had the clean individual fight stats cleaned from our main dataset, we were able to obtain aggregate stats for each fighter. This was an uncomplicated process of grouping red and blue fighters, combining them together, and getting the average of each statistic.

Lastly, we cleaned our secondary dataset of the pay-per-view data. We first had to combine the year, month, and date columns into one date column to easily match with our other datasets. While attempting to join this dataset with our main dataset, we noticed and manually fixed some inaccuracies found in the dataset with the UFC fight dates. Finally, we standardized the fighters in the dataset with the fighters from our main dataset because they sometimes used a fighter's nickname instead of the fighter's last name.

Forthcoming

With our data cleaned and in a usable format, we feel that the heavy lifting is complete, and we can now enter the next step of our process and explore the data by using different types of machine learning algorithms. This will give us initial insights into the data to understand what factors play a role in answering our questions of. As mentioned in our Team Process section, we plan to include models such as decision trees, random forest, logistic regression, and linear regression. We also have the bonus of being able to reference the literature we reviewed for ideas of ways to try different models and factors. It will be interesting to see where the team lands in answering our "Who will win?" question in comparison to the literature. We are expecting to identify factors that play a significant role in answering that question and have a more accurate model than McKinley McQuaide's conclusion of predicting the red fighter ("Applying Machine Learning Algorithms to Predict UFC Fight Outcomes."). To answer, "How to sell more pay-per-views?", we are expecting to see that fighters on a win streak that do not go to decision to be among the fighters that sell higher amounts.

Works Cited

"Topic: Ultimate Fighting Championship (UFC)." Statista, Statista Research Department, 10 Jan. 2024, www.statista.com/topics/3376/ultimate-fighting-championship-ufc/.

Tighe, D. "World's Most Valuable Sports Business Brands 2019." Statista, 22 Feb. 2024, www.statista.com/statistics/253349/brand-value-of-sports-businesses-worldwide/.

Published by Statista Research Department, and Dec 8. "Most Attended UFC Events 2022." *Statista,* 8 Dec. 2022, www.statista.com/statistics/682422/most-attended-ultimate-fighting-championship-events/.

Bashir, Umair. "Interest in Sport Types in the U.S. 2023." Statista, 13 Feb. 2024, www.statista.com/forecasts/1388963/interest-in-sport-types-in-the-us.

McQuaide, McKinley. "Applying Machine Learning Algorithms to Predict UFC Fight Outcomes." Standford. Edu, 2019, https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647731.pdf.

drwismer. "Predicting UFC Fight Scoring - Multivariate Linear Regression." GitHub, 9 June 2021,https://github.com/drwismer/metis_regression_module/blob/main/project_writeup_ufc.md#predicting-ufc-fight-scoring---multivariate-linear-regression.

Bartoš, Mikoláš. "Machine Learning in Combat Sports." Czech Technical University in Prague, 10 Aug. 2021, https://dspace.cvut.cz/bitstream/handle/10467/96672/F3-BP-2021-Bartos-Mikolas-machine_learning_in_combat_sports.pdf.