Predicting Mixed Martial Arts Match Results Through Machine Learning

Jason Oh
Adviser: Xiaoyan Li

Abstract

This study uses metrics available on `http://ufcstats.com` *to generate ML solutions for two Mixed Martial Arts prediction tasks: The "Posterior Prediction Task" and the "Prior Forecasting Task". The Posterior Prediction task is equivalent to making a fight outcome prediction between the end of the match and the reveal of the judges' decision. The Prior Forecasting task is equivalent to making a fight outcome prediction before the match has started. Logistic Regression is able to achieve a predictive accuracy of* 84.4% *for the Posterior Prediction task while models engineered for the Prior Forecasting task only achieve accuracies of around* 64%*, which is equivalent to the baseline accuracy of predicting that the red fighter wins for each match.*

1. Introduction and Problem Background

1.1. Why Should MMA be Worthy of ML Research?

The sport of Mixed Martial Arts (MMA) has evolved dramatically since its initial reputation as a crude and barbaric sport when the first UFC event debuted in 1993. It is now a well established sport, organized under many institutions internationally such as the UFC, Bellator MMA, and ONE Championship. According to Nielson Sports DNA, MMA has 451 million fans worldwide, making its popularity comparable to soccer, which has 901 million fans. [7] The growth of MMA has also led to the establishment of a global community of enthusiasts that connects through social media and online platforms. Additionally, the market for consumer MMA gyms has been growing worldwide, with 2.34 million MMA training participants in the U.S. alone, [8] encouraging a culture of fitness and empowerment for both men and women. Although there is a significant focus on the economic incentives of researching MMA, such as the engineering of better prediction algorithms in the context of sports betting, I believe that the social benefit of technically advancing MMA is a strong enough motivator for research in its own right.

My research investigates the task of predicting the outcome of a hypothetical MMA match between 2 fighters. There are many stakeholders who would benefit from a reliable match prediction model. In particular, such a model has the potential to help MMA promotion organizations (e.g. the UFC) make more balanced and thus compelling matchups between fighters. Additionally, with better prediction tools, MMA coaches and fighters can make more informed decisions regarding their training and in game strategies. Finally, the democratization of a reliable match prediction model has the potential to spark more informed discussion within the community of MMA enthusiasts, thus augmenting the general level of engagement.

1.2. The State of ML Research for MMA

MMA is a strategically rich sport, requiring competitors to be well versed in the key technical domains of striking, wrestling, and grappling. Fighters, by virtue of coming from diverse backgrounds, bring combat knowledge belonging to a highly varied spectrum. For instance, many fighters attribute their combat fundamentals to Muay Thai, a highly effective striking discipline, while others may focus on building a sophisticated library of Jiu Jitsu techniques, which eliminates striking in favor of holding and controlling your opponent from the ground. Despite MMA's evident strategic asymmetry and depth, research literature regarding the utilization of machine learning (ML) to model the dynamics of MMA is quite limited.

Thus, MMA stands in contrast to mainstream sports such as soccer and baseball, which enjoy a relatively wealthier body of research that demonstrates the successful application of ML techniques to match prediction analytics. For instance, a rigorous approach to baseball analytics originates in the 1970s when Bill James proposed *sabermetrics* [11]. Accordingly, websites such as FanGraphs provide such statistics which studies have used to generate predictive ML models via methods ranging from support vector machines (SVMs), artificial neural networks (ANNs), and convolutional neural networks (CNNs) [11]. In comparison, no thorougly rigorous statistical models have been published as of date regarding the prediction of UFC fights [6]. Additionally, research related to MMA match prediction is limited to resources provided by `http://ufcstats.com`, which provides a comprehensive collection of information regarding all UFC fights to date, although the provided statistical features are limited considering MMA's technical complexity.

2. Related Work

2.1. Holmes, McHale, and Zychaluk

The researchers from "A Markov chain model for forecasting results of mixed martial arts contests" (2022) [10] engineered a markov chain model, condensing an MMA bout into key approximate states, using data from 2001 to 2018 that was scraped from `http://ufcstats.com` and `https://espn.com`. The collected dataset amassed metrics from $4,678$ fights and $1,680$ unique athletes. The features engineered for inclusion in the markov model can be roughly divided into "work rate models" and "accuracy models". Namely, the volume of strikes, takedowns, and submissions attempted by a fighter were modeled as Poisson distributions while the accuracies of strikes, takedowns and submissions were modeled as Binomial distributions. These models were then incorporated as transition probabilties in the engineered markov chains that generalize MMA match states (e.g. $\{\text{standing}, \text{ground}, \text{standing strike attempt}, \text{standing takedown attempt}\} \in$ state space). Through running $10,000$ simulations per fight, the researchers were able to achieve a prediction accuracy of 61.77%. For comparison, a Bradley-Terry model (accuracy: 54.13%) and logistic regression model (accuracy: 47.71%) were also trained. The approach to evaluating the models' performance was to train using data from 2001 to 2017, then evaluate on data from 2018. With logistic regression, the statistically significant parameters ($p < 0.01$) were found to be "Strikes landed per second", "Takedowns attempted per second", "Strike defense", and "Takedown defense".

2.2. Hitkul, Aggarwal, Yadal, and Dwivedy

The researchers from "A Comparative Study of Machine Learning Algorithms for Prior Prediction of UFC Fights" (2018) [6] conducted a comparative analysis of various ML algorithms applied to MMA match forecasting, using data from http://ufcstats.com. Notably, the generated models are truly forecasting given that a time dependent dataset was constructed, reflecting what each fighter's statistics were prior to a fight. While there is no indication that, barring the time dependent manipulations, the researchers engineered new metrics (e.g. f: (total # attempted strikes, total # landed strikes, seconds) $\rightarrow$ (strike accuracy, strike rate) from the previous study), a performance comparison was made against the following approach: halving the number of features by taking the ratio of fighter A's and fighter's B's metrics (e.g. f: (# strikes A, # strikes B) $\rightarrow$ # strikes A / # strikes B). The assembly of ML algorithms used were: KNN, Decision Tree, SGD classifier, Random Forests, SVM, Naive Bayes, and Perceptron. The researchers found that Random Forests (accuracy: 61.15%) and SVM (accuracy: 59.80%) provided the best results in terms of prediction accuracy. Notably, Naive Bayes suffered the worst accuracy of 21.28%. The benefits of collapsing fighter A's and fighter B's metrics into a single ratio are inconsistent. Namely, while there is a $+1\%$ increase in accuracy with the use of ratio features for Random Forests, SVM, and Perceptron, there is a -1% decrease in accuracy for KNN, Decision Tree, and SGD classifier. This inconsistency suggests that there is little benefit to providing the models with ratio features.

2.3. Robles and Wu

Robles and Wu (2015) [12] also conducted a comparative analysis of various ML algorithms applied to MMA match prediction, using data from http://ufcstats.com. They collected 217 training samples and tested on 58 samples, generating the following prediction accuracies: Linear-Kernel SVM: 68.1%, Logistic Regression: 69.82%, and Naive Bayes: 56.89%. Many SVM kernels were tested, with significant variations in performance, and the Linear kernel demonstrated the highest prediction accuracy. Similar to the study by Hitkul et al., fighter A's and B's features are collapsed to the ratios of features (f: (A, B) $\rightarrow$ A / B). Additionally, 8 features were engineered for each fighter, including “Significant Strikes Landed per minute”, “Significant Strike Accuracy”, “Significant Strikes Absorbed per Minute”, “Significant Strikes Defense”. It is important to note that unlike the previous studies by Holmes et al. and Hitkul et al., these models make predictions using statistical match posteriors rather than exclusively match priors. To clarify, the model utilizes the match’s summary statistics to generate a prediction although it is only possible to possess such statistics after the match has ended, making Robles and Wu’s benchmark prediction scores incomparable to those achieved by the aforementioned studies that make “true” forecasts.

3. Approach

3.1. Overview of ML Model Selection

Logistic Regression models the probability of an event taking place as the log odds of a linear combination of dependent variables. The coefficients of the linear combination (i.e. $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$) are fit via optimization techniques such as gradient descent to best explain the training data. While logistic regression is perhaps one the most modest ML frameworks, it is still a fundamental tool and often all that is needed to generate powerful baseline results. Considering that the overall goal here is to generate various “proof of concepts” rather than demonstrate a final fine tuned product, Logistic Regression will be an especially valuable friend to this study. [2]

A Support Vector Machine is an ML model that belongs to the family of linear classifiers such as Logistic Regression. At a high level, SVMs aim to separate data points with a linear plane, like Logistic Regression, but in a way such that the distance from the plane to the nearest data point on each side is maximized. Notably, the performance of an SVM depends on the selection of its hyperparameters such the kernel (function that maps data to higher dimensional space). In practice, a popular kernel choice is the “Radial Basis Function”, and this is the default kernel in sklearn’s SVC (Support Vector Classification) library. [4]

A Decision Tree is another ML paradigm that is modest yet powerful (like Logistic Regression), modeling the decision task as a tree like structure of decisions. The decision making path in a trained tree starts from the root node and ends at a leaf node, yielding the classification prediction.

[1] Although this study does not interpret its trained decision trees, doing so could yield useful insights. As an extension of the Decision Tree paradigm, Random Forests is an ML learning method that constructs a collection of decision trees at training time, outputting the majority vote of these trees when executing a classification task. Random Forests generally perform better the use of a single decision tree, but lacks its simple paradigm and explainability. [1] [3]

We can apriori suppose in this study that the engineered features likely are not independent from each other. [2] Since Naive Bayes classifiers make a strong independence assumption between the input features, it makes sense for this study to not select Naive Bayes. Additionally, boosting methods and ANNs (artificial neural networks) are not considered in order to limit the study's scope.

3.2. Data Collection

This study scrapes UFC match statistics from `http://ufcstats.com`. Figures 17, 18, 19 show the general layout of the website. The main page lists the history of all completed UFC events up to date. Clicking on a UFC card (i.e. event) gives information as to what matches took place. Each UFC card contains a list of 10 or more matches that took place that day. Of these matches, one is designated as the "title" fight, consisting of five 5-minute rounds at maximum, while all other matches consist of three 5-minute rounds at maximum. Finally, clicking on any match provides cumulative and round-by-round metrics regarding the fighters' overall striking, wrestling, and grappling performances. Additional details about the match, particularly the number of rounds and time elapsed, as well as the match end method are provided as well.

3.3. Feature Engineering

Figure 20 details the metrics that this study collects from the fight details pages of `http://ufcstats.com`. There are some notes to be made. First, the only striking metrics that are utilized further in this study are "significant strikes" and "all strikes". Thus, metrics on head, body, leg, distance, clinch, and ground strikes are not used although they remain available in the CSV

[1] this is not a problem for us as only Logistic Regression is interpreted in this study.
[2] e.g. strike rate and strike accuracy go hand in hand.

file generated by data collection. This decision was made purely out of convenience; thus future work may wish to evaluate the effects of including these features. Figure 1 details the features that were engineered from the metrics provided by http://ufcstats.com. This study's approach to feature engineering is quite simple and falls into two general categories. First, given a certain attack such as "strikes", the number of attempted strikes and the number of landed strikes are converted into an accuracy metric (i.e. number of attempted strikes / number of landed strikes). Second, given a certain attack such as "strikes" again, the number of strikes and the match duration are converted into a rate metric (i.e. number of attempted (or landed) strikes / match duration). These are simple yet important manipulations. For instance, the number of attempted strikes is uninformative by itself. It is unknown whether the majority of these attempted strikes missed and whether these attempts are a lot relative to the total match duration, hence the importance of engineering features.

Grappling Metrics	
	Elaboration
Control Time Ratio	Percent of the total duration of the match that the fighter is able to hold his opponent on the ground from a position that is favorable for further offensive action.
Takedown Success Ratio	Takedowns are throws and tackles that a fighter attempts in order to transition the state of the match from standing to wrestling on the ground.
Takedown Attempt Rate	Number of takedowns attempted per minute.
Submission Attempt Rate	A fighter with ground control can make a submission attempt (e.g. chokes, joint locks). A successful submission attempt yields a fighter with a victory.
Relevant Martial Arts	Brazilian Jiu Jitsu, Wrestling, Judo, Sambo, etc.
Striking Metrics	
	Elaboration
Strike Hit Ratio	Strikes include a diverse range of attacks including punches, kicks, knees, and elbows. Each has a variety of techniques.
Strike Attempt Rate	Number of strikes attempted per minute.
Strike Hit Rate	Number of strikes landed per minute.
Relevant Martial Arts	Muay Thai, Boxing, Karate, Kickboxing, etc.

Figure 1: Elaboration on Engineered Features.

3.4. The Posterior Prediction Task

In the posterior prediction task, we are allowed to use the retrospective metrics from a fight in order to generate a prediction. In other words, we are allowed to use any information from the http://ufcstats.com "fight details" pages except for the actual winner to generate a prediction.

It may seem that a solution to the posterior prediction task would be quite pointless; a fighter can't adjust his strategy based on a model's prediction once his match has concluded. Moreover, a sports better can't switch his call once the winner has been revealed. Nevertheless, there are important applications for effective posterior prediction models. First, consider that a little less than half of all UFC fights have ended via judges' decisions as shown in Figure 6. With this in mind, a model that can "learn" and generalize the judges' decision making process, evaluating the fighters' performances [3] to reach a verdict is valuable. First, it can serve as an algorithmic "judge", both consulting and auditing the decision making panel by generating its own prediction. Moreover, the outcome generated by the algorithmic "judge" can be presented to the audience between the end of the match and when the judges announce their decision, in order to inform the audience and improve their viewing experience. Second, the posterior prediction model can serve as a self evaluation tool for fighters. For instance, there is sometimes the concern that the outcome of a match was "lucky" (e.g. ended via a miracle kick), or that it doesn't quite reflect the fighter's performance. The model can lend insight as to whether their performance is correlated with the fight outcome that was realized.

3.5. The Prior Forecasting Task

In the prior forecasting task, we are only allowed to use the metrics that are already available before the start of a fight in order to generate a prediction. In other words, we are truly forecasting akin to sports better who places their call before the the fight starts. A successful prior forecasting model has many important applications. One such application is generating a tool that allows fighters to anticipate match results in order to inform their training methods and fight strategy. Another is allowing MMA promoters to organize the most interesting fights by identifying pairs of fighters with equal odds of winning. Unfortunately, `http://ufcstats.com` does not conveniently provide any metrics for conducting the prior forecasting task. `ufcstats.com/fighter-details` does detail fighters' up to date career statistics as shown in Figure 2, but it can not be used to

[3] hopefully they're not taking anything else other than the fighters' performances into consideration (e.g. ethnicity, height, etc). Examining MMA judges' decisions for potential bias could be a topic for future work.

Figure 2: Khabib's Career Statistics as of 4/30/23.

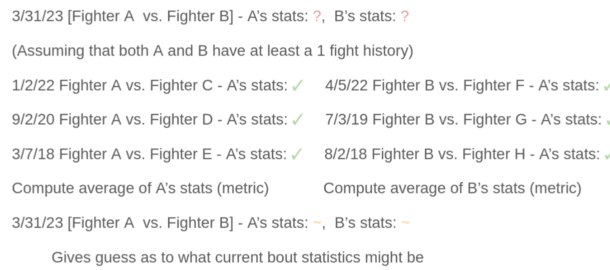


Figure 3: Mean Imputation Schema.

generate predictions for the figher's history of matches in the experimental setup because we are leaking future metrics to past events. [4] Therefore, this study generates its own time series of career statistics for each match, respecting the need to keep information from future matches away from past ones. Figure 3 illustrates the schema to engineer forecasting metrics for each match. There are multiple approaches to engineering forecasting metrics. The two simplest approaches are either to take the average of a metric throughout a fighter's entire career history or to only take the metric from their most recent fight. Alternatively, another approach, among many others, is taking a weighted average that favors information from a figher's most recent fights rather than their old ones. This strategy is definitely worth considering in future works, but this study only considers the aforementioned two basic methods for engineering forecasting metrics.

4. Implementation

Work for all parts of this project was done in Google Colab, a cloud based Jupyter notebook environment for Python programming that runs on your web browser. The Colab environment was also linked my Google Drive system where intermediate results were stored. For the web scraping portion of the project, Python's MechanicalSoup and Pandas libraries were utilized. The MechanicalSoup library simulates the functionality of a web browser, allowing you to automate your interactions with websites. MechanicalSoup itself imports the BeautifulSoup library as a dependency, which provides functionality for parsing and extracting data from HTML documents. MechanicalSoup as well as its BeautifulSoup dependency made it possible for me to collect metrics

[4] nevertheless, it is probably safe to make an exception for the height and reach metrics (but not weight and stance).

from thousands of UFC matches. Additionally, scraping results were collected in the format of a Pandas dataframe before they were exported to XLS format.

Moving on, feature engineering and a myriad of other data manipulation tasks were done via Pandas functionality. Additionaly, Python's Matplotlib library was used to generate various visualizations of data when I was trying to understand it. Finally, training, testing, and evaluation on machine learning models was done via the tools provided by the Python scikit-learn library. In particular, scikit-learn provided a very pleasant interface for importing the various ML models used in this study and interacting with them. Here I provide a link to my project GitHub repository: `https://github.com/Jason0h/MMA-Match-Prediction`.

5. Evaluation

5.1. Findings From Data Exploration

The examination of summary statistics illuminates various basic yet crucial findings regarding overall trends in the MMA match datasets. To begin, the raw dataset contains the metrics of $6,987$ total matches that were held over the roughly 30 year period from UFC 2 ($03/1994$) to UFC Fight Night: Muniz vs Allen ($02/2023$). Anomalous matches were eliminated from this dataset (e.g. no contest, missing data, etc). Luckily, with the elimination of such matches, $6,716$ were left remaining in the dataset, ranging from the dates $07/99$ to $02/23$. For comparison, Holmes et al. amassed statistics from $4,678$ fights (Section 2.1), providing my study with a significantly greater advantage in terms of data volume. Continuing on, 172 of the 271 matches that were eliminated originate from the founding era of the UFC (1993 - 1999). This is reasonable given that UFC 21 ($07/1999$) marks the introduction of multi-round fights and the move to standardize MMA's ruleset. [9] Thus, the filtered raw dataset can be assumed to be mostly homogenous in terms of the general match ruleset, with the majority of fights consisting of 3 rounds and title bouts consisting of 5 rounds. [5]

Figure 4 shows the distribution of $5,700$ men's fights over the 8 standard UFC men's weight divisions, ordered by increasing weight classes. It roughly follows a normal distribution, with

[5] an exception is that some 2 round matches persisted until 2001.

most fighters belonging to the Lightweight and Welterweight classes. On the other hand, Figure 5 shows the distribution of 587 women's fights over the 4 standard women's weight divisions, ordered by increasing weight classes. The main takeaway here is that the number of men's UFC matches outnumbers that of women's UFC matches by 10 : 1. Given that the volume of training data is skewed significantly towards men's matches, it is worth comparing the predictive models' performances on men's and women's matches, on the principle of algorithmic fairness.

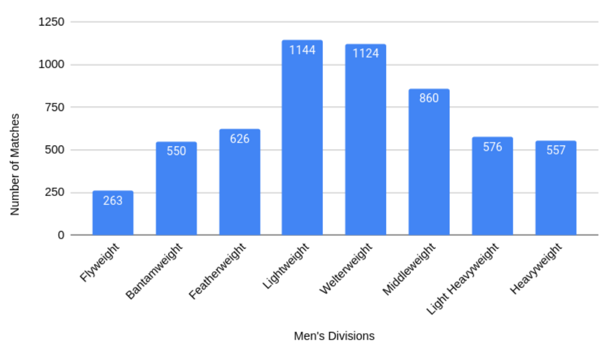


Figure 4: Matches among Men's Divisions.

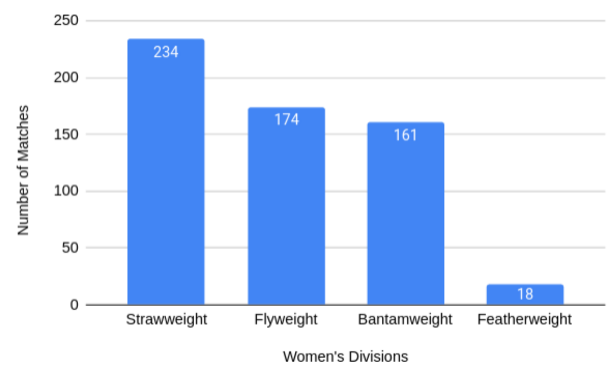


Figure 5: Matches among Women's Divisions.

Figure 6 shows the distribution of match end methods. KO/TKO generally means that the match ended via one of the fighters incapacitating the other through the use of strikes (punches, kicks, etc). Submission means that the match ended via one of the fighters forcing the other to yield with either a choke or joint lock. Finally, Decision means that the judges determined the winner of the match, if it was the case that neither fighter was able to knock out or submit the other. The distinction between matches that end via KO/TKO/Submission and Decision is very important. Namely, the 52.9% of fights that do not end via decision are susceptible to "lucky" outcomes. To elaborate, a fighter can reverse the outcome of a match at any time via a "lucky" knockout or submission, even if relevant metrics and eyewitness observation undoubtedly show that the fighter was performing poorly overall up to that point. On the other hands, decisions are not "lucky" in that the judges make a holistic evaluation of the fighters' performances throughout the entire match.

The standard outcomes of an MMA match are the following: red win, blue win, and draw.

Rather than making the problem a task of multiclass classification, I focused on binary outcomes, where the opposite of red winning is (red losing + red drawing with blue). [6] To further justify this decision, only 52 matches from the filtered dataset result in a draw (0.6%). Thus, draws are not significant enough to this study for them to be added as an additional classification outcome. Overall, 64.7% of the matches ended with the red fighter as the winner. This raises an important point. If our strategy is to forecast that red wins the match before every fight, the expectation is that we'd be right 64.7% of the time. While the previous body of related work fails to provide a fundamental benchmark to evaluate model performance, it will be the case in this study that models achieving a classification accuracy below 64.7% are deemed unsuccessful.

Figure 7 shows the red and blue fighters' average performances among the cohort of features engineered for the study. It is remarkable that the red fighter, on average, demonstrates a higher level of performance on all engineered metrics (attack accuracy, attack frequency, time spent ground controlling opponent). This finding correlates well with the earlier finding that 64.7% of matches end with the red fighter as the winner. Additionally, to add more background, fighters from the red corners are usually the favorites (veterans), while those from the blue corners are usually the contendors (underdogs). In theory, it is to be expected that the favorites are better fighters that the contendors. [5] Thus, even the corner color alone is a valuable predictive variable.

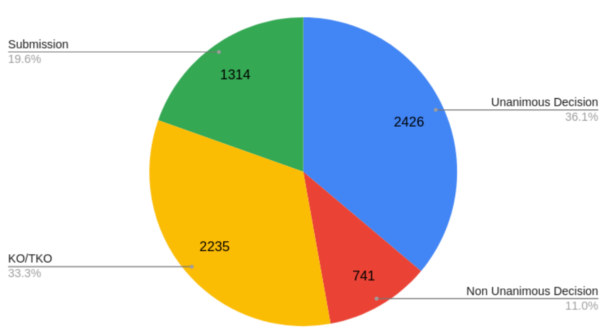


Figure 6: Distribution of Match End Methods.

	Red		Blue	
	mean	std	mean	std
Percent Match Completed	67.4%	36.2%	67.4%	36.2%
All Strikes Hit Ratio	55.5%	17.3%	51.6%	18.3%
All Strikes Attempt Rate	10.50	5.90	9.36	5.63
Sig Strikes Hit Ratio	47.9%	16.2%	43.7%	17.2%
Sig Strikes Attempt Rate	8.48	5.79	7.75	5.50
Takedown Success Ratio	62.5%	39.6%	59.6%	41.9%
Takedown Attempt Rate	0.289	0.357	0.259	0.364
Submission Attempt Rate	0.0761	0.223	0.0455	0.145
Percent Match in Control	24.6%	25.3%	17.5%	20.9%

Figure 7: Statistics From Basic Feature Eng.

[6]likewise, the opposite of blue winning is (blue losing + blue drawing with red). For simplicity, my study only tackles the red win / red not win classification task.

5.2. Evaluation of The Posterior Prediction Task

First, the study evaluated the predictive value of different combinations of engineered features. Figure 8 illustrates the results of this step. Logistic Regression was used for training and testing on all parts in this table to control for ML algorithm selection. The engineered features can be roughly divided into 2 categories: grappling metrics and striking metrics. Grappling metrics indicate a fighter's ability and rate with which they are able to 1. bring the opponent to the ground, 2. keep the opponent on the ground, 3. finish the fight from the ground. Striking metrics indicate a fighter's ability and rate with which they are able to execute striking attacks. Figure 1 once again elaborates further regarding these features.

	Train Accuracy	Test Accuracy	Precision	Recall
Grappling	75.1%	75.4%	77.1%	88.0%
(All) Striking	79.8%	77.9%	78.8%	90.0%
(Sig) Striking	80.7%	79.4%	80.2%	90.4%
(All + Sig) Striking	81.4%	80.4%	81.1%	90.9%
(Sig) Striking + Grappling	85.4%	86.0%	87.0%	91.9%

Figure 8: Logistic Regression Performance on Various Features.

	Red	Blue		Statistical Significance
Control Time Ratio	3.9e-22	2.0e-10		
Takedown Success Ratio	4.0e-09	1.5e-04		
Takedown Attempt Rate	2.3e-01	5.8e-02		Very Strong
Submission Attempt Rate	3.4e-45	2.2e-35		Strong
Sig Strike Hit Ratio	2.6e-06	2.4e-03		Weak
Sig Strike Attempt Rate	1.2e-03	3.9e-01		
Sig Strike Hit Rate	1.3e-07	9.9e-18		

Figure 9: P-Values of Engineered Features.

Predicting the match outcome using only striking metrics, with a best accuracy of 79.4% is more effective than predicting the match outcome using only grappling metrics, with an accuracy of 75.4%. Nevertheless, both approaches perform well above the baseline classification accuracy of 64.7% that we've established from before. http://ufcstats.com provides metrics on "all strikes" and "significant strikes", and significant strikes are a strict subset of all strikes. Although the prediction accuracy increases by 1.0% (from "significant strikes") when "all strikes" + "significant strikes" metrics are both included, the improvement is negligible enough that only significant strike metrics are utilized from this point on. [7] Finally, when striking and grappling metrics are combined (14 features in total: 7 from red, 7 from blue), a classification accuracy of 86.0% is achieved, indicating that the combined use of both striking and grappling metrics is important.

It is quite remarkable that recall is 88% when using grappling metrics only, improving up to

[7] figure 21 also shows that significant strikes and all strikes metrics are highly correlated.

91.9%, when using (Sig) Striking + Grappling metrics. Recall is defined as True Positives / (True Positives + False Negatives). In other words, suppose we know that a match ended with the red fighter as the winner; recall measures the ability for a model to then predict that red did indeed win, that is to say the Logistic Regression models have impressive abilities to do so. Precision is defined as True Positives / (True Positives + False Positives). In other words, it is the accuracy of a model's predictions that the red fighter wins. The precision scores roughly mirror the accuracy scores.

Figure 9 shows the p-values of features from the model trained on (sig) strikes + grappling metrics. For the most part, most of the engineered features have a very strong level ($p < 0.001$) of statistical significance, meaning that they are significantly associated with the match outcome. The exceptions are "takedown attempt rate" and "significant strike attempt rate". This suggests that attempt rates are not quite statistically significant when predicting match outcome. For instance, a fighter might throw a lot of significant punches per minute throughout the match. However, it remains ambiguous whether these are well timed and calculated punches, or punches thrown in a haphazard manner in the hope that one of them lands. On the other hand, significant strike hit rate, or the rate at which successful punches are thrown, has a much greater deal of statistical significance because we can assume that these punches are well timed and accurage on average. A question that might follow is why "submission attempt rate" has a high statistical significance. Unlike takedown and strike attempts, submission attempts are most often made when the fighter has achieved a favorable position from ground control. In other words, in general, the chance to make a submission attempt is "earned" through having demonstrated superior grappling skills.

Figure 10 illustrates the evaluation metrics of a menagerie of ML models that utilize the aforementioned (Sig) Striking + Grappling features. There are some important comments to be made. Namely, Logistic Regression, Support Vector Machine, and Random Forests were overall able to achieve prediction accuracies of around 84%, which is roughly 20% better than the baseline strategy of predicting that the red fighter wins for each match. Among the main classes of classifiers, Logistic Regression, SVM, and Random Forests achieve a best classification accuracy of 84.4%, 84.7%, and 85.6% respectively. Hyperparameter selection, via regularization and kernel choice for Logistic Regression and SVM respectively does not appear to be particularly important for test accuracy. Nevertheless, it may be of some importance that choosing the Radial Basis Kernel for SVM yields a recall score of 94.6%, which is 3% greater than the next highest recall score of 91.6% from Random Forests. [8] Finally, it is significantly more effective to use Random Forests than Decision Tree (85.6% accuracy vs. 77.9% accuracy).

	Train Accuracy	Test Accuracy	Precision	Recall
Predict Red Wins	----------	64.4%	64.4%	100%
Logistic Regression	85.6%	84.4%	86.2%	90.3%
Logistic (L1 Reg)	85.7%	84.4%	86.2%	90.3%
Logistic (L2 Reg)	85.6%	84.4%	86.2%	90.3%
SVM (Linear Kernel)	85.6%	84.2%	85.7%	90.4%
SVM (Cubic Kernel)	83.1%	83.3%	82.2%	94.6%
SVM (Radial Basis Kernel)	85.1%	84.7%	85.9%	91.2%
Decision Tree	78.4%	77.9%	82.6%	83.2%
Random Forests	84.9%	85.6%	86.8%	91.6%

Figure 10: Evaluation of Model Selection on (Sig) Striking + Grappling Features.

	Train Accuracy	Test Accuracy	Precision	Recall
Predict Red Wins	----------	57.9%	57.9%	100%
Logistic Regression	85.6%	84.1%	83.3%	90.7%
Logistic (L1 Reg)	85.6%	84.0%	83.3%	90.5%
Logistic (L2 Reg)	85.6%	84.1%	83.3%	90.7%
SVM (Linear Kernel)	85.3%	83.9%	83.1%	90.7%
SVM (Cubic Kernel)	82.4%	82.4%	78.6%	95.7%
SVM (Radial Basis Kernel)	85.1%	84.1%	83.1%	91.0%
Decision Tree	78.9%	77.1%	78.7%	82.9%
Random Forests	85.3%	83.8%	83.4%	89.9%

Figure 11: Evaluation When Test Dates > 1/7/20 and Train Dates < 1/7/20.

Here in Figure 11, we see that the study violates standard test/train split procedure. The training set consists of $5,366$ fights including and before the $6/27/20$ UFC card. The testing set consists of $1,350$ fights including and after the $7/11/20$ UFC card. This provides us with a roughly $20:80$ test/train split that has been used in all the previous evaluations. However, this test/train split is obviously not randomized unlike the previous ones. Our "temporal" split is important for evaluation because it reflects the real world application of our model. With a random split, it is always the case that there exists a match from the training set whose card date is greater than the card date of a match from the testing set. [9] In other words, our models have been using "future data" to generate its predictions, which is impossible. [10] The "temporal" split solves this concern. Like the previous set of models, the discrepancy between training accuracy [11] and testing accuracy is minimal ($< 2\%$), meaning that past data generalizes very well to future matches. Additionally, test

[8] actually, the highest recall score is 100% from the "predict red every time" strategy :)
[9] of course, random splits that also happen to also be temporal splits are rare exceptions :)
[10] in principle... but theoretical time travel is outside the scope of this study :)
[11] training accuracy is calculated via 5-fold cross validation on the training set in this study.

accuracy from the "temporal" split here is consistent with test accuracy from the "random" split before. On the other hand, precision drops consistently by around 3% among all models. A final note is that the red wins 57.9% of the time from $7/11/20-$ and 66.4% of the time from $-6/27/20$. Figure 12 further explores this phenomenon. The chart shows the percent of matches where red wins among 5 roughly evenly spaced periods (roughly 1,340 matches each). Remarkably, the red fighters have historically won 91% of fights on average from 1999 to 2011. This is in stark contrast to the fact that red fighters have won around 58% of matches from 2011 to 2023. [12] This suggests that in practice, 58% is the expected accuracy for the "predict red win" strategy, not 64%. The phenomenon may be explained in part by the "modernization" of MMA. To elaborate, MMA fighters today are as a whole much better prepared than they were in the past because of the lessons that have been learned. For instance, it is less likely today for Jiu Jitsu specialists to completely neutralize strikers' offensive capabilities or Muay Thai champions to dominate matches with leg kicks; modern MMA fighters bring well rounded offensive + defensive knowledge and training in striking, grappling, and wrestling before entering the ring. Nevertheless, this is merely a hypothesis. Further domain knowledge and statistical experiments would definitely be needed to present a convincing case for this argument. Finally, it may be worthwhile to compare model performance when training/testing on 2011 - 2023 match data exclusively, but that is unfortunately left to future work.

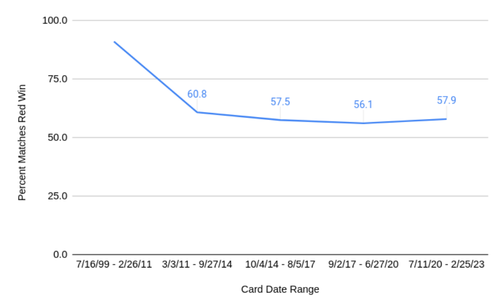


Figure 12: Evolution of Matches Won by Red.

	Accuracy	Precision	Recall	Test Size
All	85.3%	85.9%	91.8%	1344
Decision	82.9%	84.2%	89.4%	637
No Decision	87.4%	87.5%	93.8%	707
Mens	85.4%	86.3%	91.7%	1151
Womens	82.1%	79.2%	91.9%	112

Figure 13: Evaluation of Logistic Regression on Different Classes.

Figure 13 evaluates the performance of Logistic Regression on different classes of matches. The testing and training sets were constructed by a 80 : 20 random split. The matches in the test set

[12] also, observe that while both periods span roughly equal amounts of time, the 2011 to 2023 period has had 4x the number of matches than the 1999 to 2011 period.

are labeled as those that did and did not end via judges' decision. Also, they are further labeled as Men's or Women's matches. [13] Consider the model's performance on matches that end via decision; in this scenario, we can say that the model is assuming the role of a judge after the match has ended. Compared to the model's overall accuracy of 85.3%, the accuracy on decision matches is 82.9%. The drop in accuracy is to be expected as decision making introduces more uncertainty. Nevertheless, this result suggests that our model shows promise for the applications of algorithmic consulting and auditing for the UFC decision making process. Moving on, the accuracy on non decision matches (i.e. matches that end via Knockout or Submission) is 87.4%, which is higher than the average accuracy. This suggests that concern of "lucky" outcomes should be lower than that which is held by popular sentiment. Overall, regression demonstrates that there is a very significant correlation between both fighters' overall performances and the outcome if it resulted via Knockout/Submission. Additionally, the prediction model can serve as algorithmic evidence for lucky outcomes if its prediction is not the same as the fighter who won via Knockout/Submission. Moving on, the Men/Women comparison suggests that there exists a slight performance bias towards Men's matches (+3.3% accuracy), though the small size of the Womens test set (112) introduces some uncertainty to the evaluation. Finally, further evaluations that future works may want to explore include: training on certain classes of matches, performance on different divisions (e.g. Men's Lightweight vs Men's Heavyweight).

5.3. Evaluation of The Prior Forecasting Task

To begin, the first consideration that had to be made regarding the prior forecasting task was whether forecasts should be made for matches where at least one of the fighters is a "newcomer". In other words, the fighter has no history of career metrics. The decision this study makes is to remove such fights from consideration, neither generating predictions from nor training models on them. Thus, while there are $6,716$ total fights in the filtered database, $4,943$ fights are left after dropping those featuring a newcomer (26.4% of fights are dropped). Moving on, Figure 14 shows the median

[13] you might have observed that 1151 Men + 112 Women = 1263 matches not 1344 matches, but it would have taken quite long to sort all unconventionally titled divisions into the Men/Women categories

number of previous matches from the career histories of the red and blue fighters. In general, red fighters have slightly longer career histories (median = 6 fights) than blue fighters (median = 4 fights). Observe that career histories range from merely 1 fight for rookies to 39 fights for battle hardened veterans. On average though, we can anticipate that fighters' career histories will be around 5 fights. From the perspective of the forecasting task, this is not quite good news because it means that there is limited precedence to generalize each fighter's anticipated performance level. On the other hand, this is not a surprising finding because fighters in MMA compete at longer intervals relative to other sports due to its physically taxing nature.

	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Red Fighter	1	3	6	10	39
Blue Fighter	1	2	4	8	36

Figure 14: Previous Number of Fights at Forecasting Time.

As described before [3.5], there are two basic approaches to anticipate a fighter's performance on their next match. Suppose that we want an estimate of a figher's striking accuracy and striking rate, and that this fighter also has a career history of 5 fights. Our estimate can be generated by averaging the fighter's striking accuracy and striking rate throughout their previous 5 fights. Alternatively, we can look at the fighter's most recent fight to generate an estimate. [14] The metrics that are used to generate prior forecasting models fall into the general categories of Meta, Offensive, and Defensive metrics. "Meta" metrics are the number of previous matches in a fighter's career history, the proportion of these matches that were won, and the estimated proportion of a match's total duration that is completed. Offensive metrics consist of the same metrics that were used in the posterior prediction task. [15] Finally, defensive metrics are identical to offensive metrics; however they measure the offensive metrics of a fighter's previous opponents. In other words, the "strike absorb rate" and "strike absorb ratio" measure how successful a fighter's previous opponents have been at throwing strikes against them. Figure 22 lists the metrics that are engineered for the prior forecasting task. There are some metrics that should have been included in the setup, but

[14] recall that weighted averages are another approach, although this study does not implement nor evaluate it.
[15] see Figure 1.

are otherwise missing. In particular, submission attempt rate and strike hit rate were omitted by mistake. Additionally, a module in the web scraper to gather fighters' height and reach information from `http://ufcstats.com` was not implemented. Future studies are highly encouraged to report potential improvements in results when these features are included.

Career Statistics : Average Metrics From Fighter's Entire History				
	Train Accuracy	**Test Accuracy**	**Precision**	**Recall**
Meta	63.4%	62.5%	62.8%	96.7%
Offensive	62.2%	63.1%	63.7%	95.7%
Defensive	64.2%	64.2%	65.5%	95.3%
Off + Def	62.2%	63.8%	64.1%	98.7%
Meta + Off + Def	62.7%	63.1%	63.9%	92.3%
Career Statistics: Metrics From Fighter's Most Recent Match				
	Train Accuracy	**Test Accuracy**	**Precision**	**Recall**
Meta	63.1%	63.9%	64.0%	97.8%
Offensive	62.6%	59.8%	60.2%	97.8%
Defensive	61.6%	64.5%	64.9%	97.3%
Off + Def	62.2%	61.4%	62.1%	97.4%
Meta + Off + Def	62.7%	64.4%	65.6%	92.5%

Figure 15: Evaluation of Logistic Regression on Prior Forecasting

Moving on, Figure 15 shows the evaluation of Logistic Regression on different combinations of features. Additionally, the evaluation is divided by the mean and last fight approaches to estimating metrics. Overall, the results are quite disappointing. None of the approaches are able to achieve a higher test accuracy than the 64.7% benchmark achieved by forecasting that the red fighter wins before each match. In fact, the test accuracies of the approaches are around 64% on average themselves. Additionally, observe that the recall scores are 96% on average; this is close to the 100% recall score that the "predict red" strategy achieves by virtue of choosing only one outcome. This should leave us to believe that the learned behaviors of the Logistic Regression models are about equivalent to that of the "predict red" strategy and no more sophisticated. Next, Figure 16, which extracts the p-values of the Meta + Off + Def metrics, shows that the overwhelming majority of features are not statistically significant in regards to the model output either. It is hard to say for sure which, if any, features can be considered as important. However, it does seem seem that the Logistic Regression models associate the Red fighter's number of previous fights to the match outcome with high confidence. Figure 23 illustrates the evaluation on different models and hyperparameters. Similar to the previous results, the testing accuracies of these models were a little bit under 64.7% as well.

	Mean Metrics		Last Match Metrics	
	Red	Blue	Red	Blue
Number of Previous Fights	1.3e-07	6.3e-02	9.6e-08	3.7e-01
Past Win Proportion	1.8e-03	1.6e-02	1.4e-05	2.9e-01
Proportion Match Completed	2.6e-02	2.1e-04	3.4e-01	3.6e-02
Sig Strike Hit Ratio	9.9e-01	9.3e-01	9.7e-01	2.1e-01
Sig Strike Attempt Rate	8.6e-01	4.9e-04	7.1e-01	2.1e-01
Takedown Success Ratio	3.93-03	5.3e-02	6.0e-01	6.9e-01
Takedown Attempt Rate	1.3e-01	9.4e-02	9.2e-02	9.4e-02
Control Time Ratio	3.1e-04	8.8e-01	2.6e-02	5.8e-01
Sig Strike Absorb Ratio	5.7e-01	6.3e-01	1.5e-02	4.0e-01
Sig Strike Absorb Rate	3.1e-02	5.3e-01	2.0e-02	3.0e-01
Controlled Time Ratio	5.6e-01	8.9e-01	6.3e-01	4.3e-03

Statistical Significance
Very Strong
Strong
Weak

Figure 16: P-Values of Features From Logistic Regression

In summary, the evaluation demonstrates that this study's approach to the prior forecasting task is ineffective. It suggests that metrics regarding fighters' previous fight performances do not generalize well to the outcomes of their future matches, at least with respect to the information that is available on http://ufcstats.com. There are many reasons for why this may be the case. First, unlike many other traditional sports such as baseball, soccer, and tennis, MMA fighters have limited career histories, 5 fights on average, which would entail less accurate estimates of metrics for future matches. Additionally, MMA fighters generally only compete around 3 times a year; this means that they have a few months to change up their strategies and physical training. Furthermore, a fighter's matches against different opponents can be vastly different from each other in terms of their overall in game dynamics. These conditions suggest that a fighter's matches could in practice be very different from each other, limiting predictive generalizability. Finally, I believe that the most critical problem is that the metrics from http://ufcstats.com are insufficient for fully capturing the mechanics of MMA. For instance, information regarding a fighter's grappling performance is limited to the time they spent in "control" and their number of reversals. However, this idea of control has many layers of complexity including but not limited to the type of control (e.g. top mount, side mount, guard, etc.) and further details regarding transitions between different Jiu Jitsu positions. In light of this, the most necessary contribution of MMA x Data Science research, with regards to pure forecasting, may be the identification of other MMA datasets or the generation of a more comprehensive database. Nevertheless future works may still wish to generate evidence against the conclusion that has been reached in this section.

6. Future Work

Ideas for future work have been proposed throughout this paper. However, I would like to discuss some that have not been introduced yet. First, evaluation of the prior forecasting task could be extended to include models that are trained and tested excusively within the subset of fights where both fighters have established career histories. Second, the success of the models trained for solving the posterior prediction tasks suggests that models trained for "live forecasting" from the metrics available on `http://ufcstats.com` would likely be successful as well. To elaborate, live forecasting generates a prediction at the end of every round, with its accuracy also growing better with the passing of each round. Such a model is valuable because it allows fighters to change their strategies and effort based on the model's anticipated outcome. Live forecasting would have naturally fit in as a third part to this project. However, this task would have required the engineering of a web scraper component that collects round by round data from `http://ufcstats.com`, which I did not implement for this project.

7. Acknowledgements

I would like to express my gratitude to my advisor, Dr. Xiaoyan Li, for her feedback and guidance throughout this project, as well as to my classmates in the Junior research seminar.

8. Honor Code

I pledge my honor that this paper represents my own work in accordance with University regulations.
Signed: Jason Oh

References

[1] "Decision tree," https://en.wikipedia.org/wiki/Decision_tree, accessed: April 28, 2023.
[2] "Logistic regression," https://en.wikipedia.org/wiki/Logistic_regression, accessed: April 28, 2023.
[3] "Random forest," https://en.wikipedia.org/wiki/Random_forest, accessed: April 28, 2023.
[4] "sklearn.svm.svc," https://scikit-learn.org, accessed: April 28, 2023.

[5] https://forums.ufcfightclub.com, 2022, accessed: April 28, 2023.
[6] H. . K. Aggarwal, N. Yadav, and M. Dwivedy, "A comparative study of machine learning algorithms for prior prediction of ufc fights," in *Advances in Intelligence Systems and Computing*, 2018.
[7] K. Badenhausen, "The new fight game: How an mma startup wants to capture the sport's 450 million global fans," https://www.forbes.com, 2018, accessed: April 28, 2023.
[8] E. Galperin, "39 martial arts industry statistics to know," https://www.gymdesk.com, 2022, accessed: April 28, 2023.
[9] J. Gugala, "How long is a ufc fight? the rules, explained," https://www.themanual.com, 2022, accessed: April 28, 2023.
[10] B. Holmes, I. G. McHale, and K. Zychaluk, "A markov chain model for forecasting results of mixed martial arts contests," in *Inernational Journal of Forecasting*, 2022.
[11] M.-L. Huang and Y.-Z. Li, "Use of machine learning and deep learning to predict the outcomes of major league baseball matches," in *Applied Sciences*, 2021.
[12] I. S. Robles and J. Wu, in *cs229.stanford.edu*, 2015.

9. Appendix

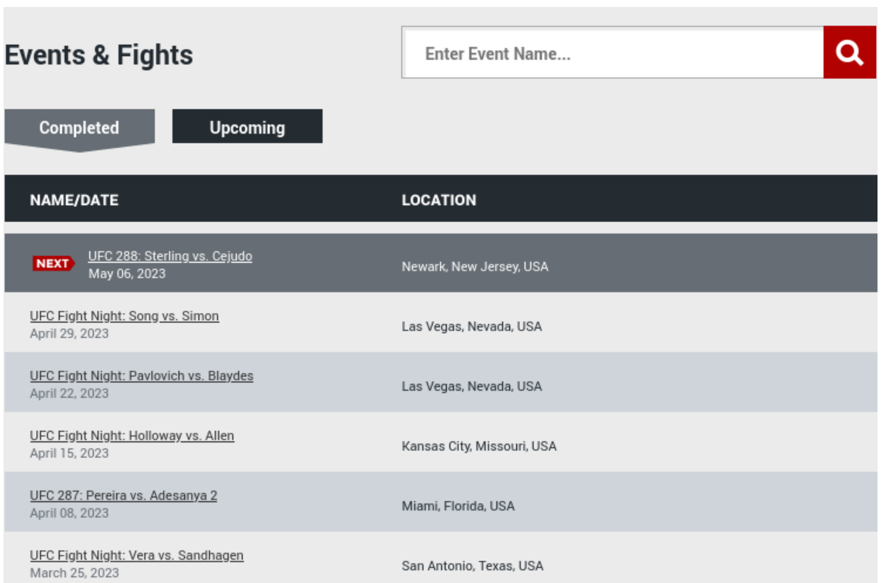


Figure 17: `http://ufcstats.com` Main Page.

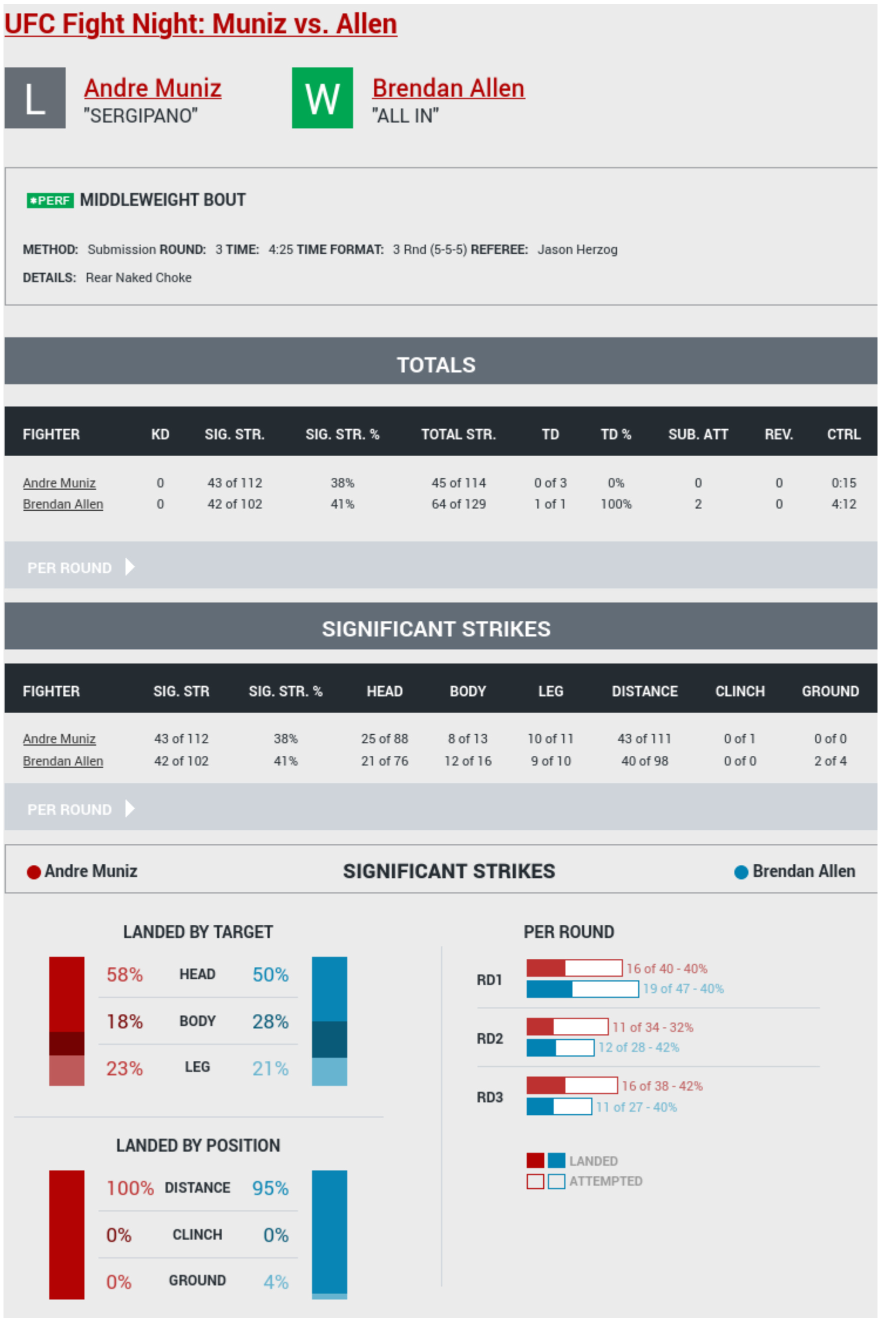
UFC Fight Night: Muniz vs. Allen

DATE: February 25, 2023 **LOCATION:** Las Vegas, Nevada, USA

Click on a row below to see in-depth event stats. **Fight, Perf, Sub, and KO of the Night Bonuses:** FIGHT PERF SUB KO

W/L	FIGHTER	KD	STR	TD	SUB	WEIGHT CLASS	METHOD	ROUND	TIME
WIN	Brendan Allen Andre Muniz	0 0	42 43	1 0	2 0	Middleweight PERF	SUB Rear Naked Choke	3	4:25
WIN	Augusto Sakai Don'Tale Mayes	0 0	53 29	1 0	0 0	Heavyweight	U-DEC	3	5:00
WIN	Tatiana Suarez Montana De La Rosa	0 0	10 5	2 0	1 0	Women's Flyweight PERF	SUB Guillotine Choke	2	2:51
WIN	Mike Malott Yohan Lainesse	0 0	9 1	1 0	1 0	Welterweight PERF	SUB Arm Triangle	1	4:15
WIN	Trevor Peek Erick Gonzalez	2 0	51 7	0 4	0 0	Lightweight PERF	KO/TKO Punch	1	4:59
WIN	Jasmine Jasudavicius Gabriella Fernandes	0 0	24 26	4 0	0 0	Women's Flyweight	U-DEC	3	5:00
WIN	Jordan Leavitt Victor Martinez	1 0	17 7	0 0	0 0	Lightweight PERF	KO/TKO Knees	1	2:27
WIN	Ode Osbourne Charles Johnson	0 0	48 60	3 1	0 0	Catch Weight	S-DEC	3	5:00
WIN	Joe Solecki Carl Deaton	0 0	2 1	2 0	3 0	Lightweight PERF	SUB Rear Naked Choke	2	4:55
WIN	Nurullo Aliev Rafael Alves	0 0	19 6	3 0	0 2	Lightweight	M-DEC	3	5:00

Figure 18: UFC Event Details.



UFC Fight Night: Muniz vs. Allen

L **Andre Muniz** "SERGIPANO"

W **Brendan Allen** "ALL IN"

PERF MIDDLEWEIGHT BOUT

METHOD: Submission **ROUND:** 3 **TIME:** 4:25 **TIME FORMAT:** 3 Rnd (5-5-5) **REFEREE:** Jason Herzog

DETAILS: Rear Naked Choke

TOTALS

FIGHTER	KD	SIG. STR.	SIG. STR. %	TOTAL STR.	TD	TD %	SUB. ATT	REV.	CTRL
Andre Muniz	0	43 of 112	38%	45 of 114	0 of 3	0%	0	0	0:15
Brendan Allen	0	42 of 102	41%	64 of 129	1 of 1	100%	2	0	4:12

PER ROUND

SIGNIFICANT STRIKES

FIGHTER	SIG. STR	SIG. STR. %	HEAD	BODY	LEG	DISTANCE	CLINCH	GROUND
Andre Muniz	43 of 112	38%	25 of 88	8 of 13	10 of 11	43 of 111	0 of 1	0 of 0
Brendan Allen	42 of 102	41%	21 of 76	12 of 16	9 of 10	40 of 98	0 of 0	2 of 4

PER ROUND

Figure 19: UFC Match Details.

Striking Metrics	Match Metrics	Grappling Metrics
Number of Significant Strikes Landed Number of Significant Strikes Attempted	Fighting Division	Number of Takedowns Landed Number of Takedowns Attempted
Number of All Strikes Landed Number of All Strikes Attempted	Match Time Elapsed	Number of Submission Attempts
Number of Head Strikes Landed Number of Head Strikes Attempted	Match Time Format	Number of Reversals
Number of Body Strikes Landed Number of Body Strikes Attempted	Number of Rounds Elapsed	Time Spent in Ground Control
Number of Leg Strikes Landed Number of Leg Strikes Attempted	Winner Name, Loser Name	
Number of Distance Strikes Landed Number of Distance Strikes Attempted	Winner Color, Loser Color	
Number of Clinch Strikes Landed Number of Clinch Strikes Attempted	Referee Name	
Number of Ground Strikes Landed Number of Ground Strikes Attempted	Match Result	
	Match End Method	

Figure 20: Basic Collected Features.

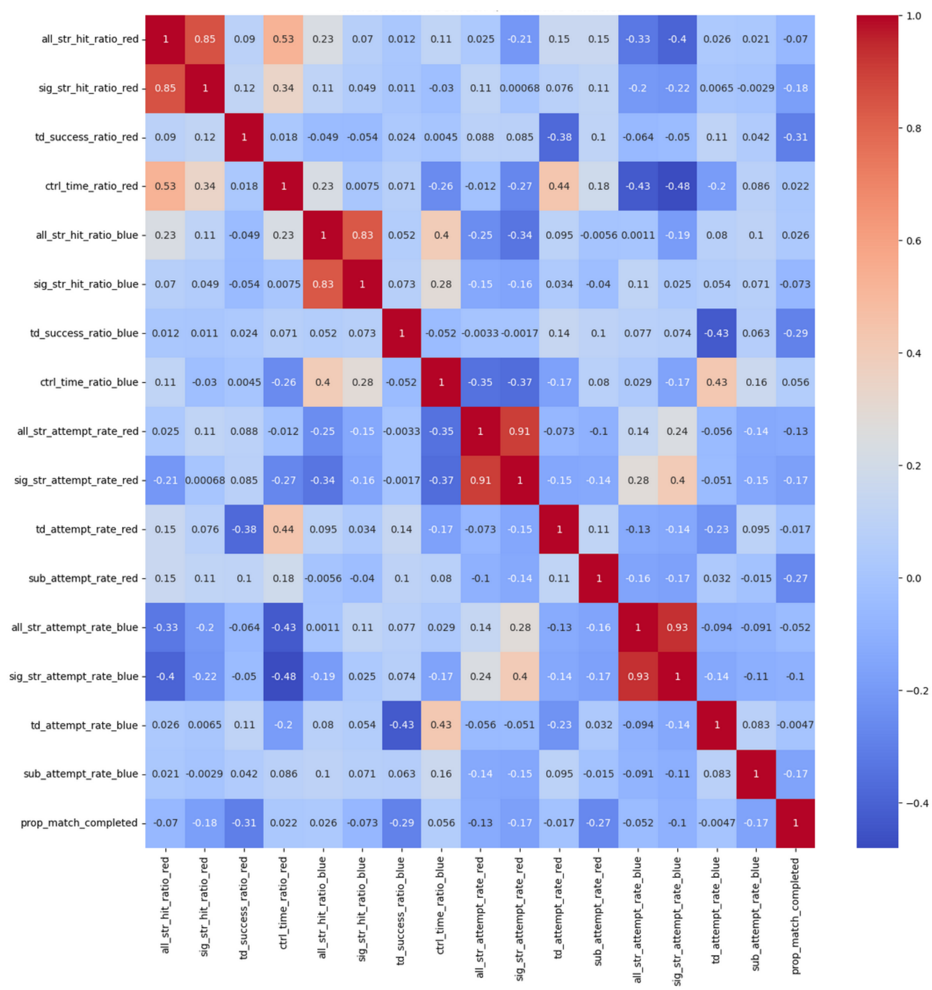


Figure 21: Correlation Between Engineered Features.

Meta Metrics	Offensive Metrics	Defensive Metrics
Number of Previous Fights	Sig Strike Hit Ratio	Sig Strike Absorb Ratio
Proportion of Fights Won	Sig Strike Attempt Rate	Sig Strike Absorb Rate
Proportion of Match Completed	Takedown Success Ratio	Controlled Time Ratio
	Takedown Attempt Rate	
	Control Time Ratio	

Figure 22: Features Used For Prior Forecasting.

	Train Accuracy	Test Accuracy	Precision	Recall
Logistic Regression	63.2%	62.9%	64.0%	91.2%
Logistic (L1 Reg)	63.3%	62.7%	63.8%	91.3%
Logistic (L2 Reg)	63.2%	62.9%	64.0%	91.2%
SVM (Linear Kernel)	62.7%	61.8%	61.8%	100%
SVM (Cubic Kernel)	61.5%	61.1%	62.6%	92.0%
SVM (Radial Basis Kernel)	63.0%	61.4%	62.3%	94.8%
Decision Tree	55.6%	55.7%	64.4%	63.3%
Random Forests	62.8%	60.8%	63.4%	86.3%

Figure 23: Evaluation of Models on Meta + Off + Def Features and Mean Career Metrics.