MGT – 6203 Group Project
Predicting UFC Fight Results
Team 55 – Sofia Laval, Will Rodriguez, Monica Singh, Greg Pawlow, Zach Powers

Introduction

The Ultimate Fighting Championship (UFC) is commonly known as the world's largest mixed martial arts organization. Yet UFC as a company is not about fighting. UFC themselves do not train fighters or have any direct influence over how fighters train or perform. UFC is about two things: marketing and matchmaking. And according to publicly available information, UFC does not have an objective means of matching two individuals. Most of the matchmaking is guesswork done by highly experienced individuals. Additionally, many of their matches are set up purely based on fighter popularity, not skill, because that is what will intuitively bring in the most money in the short term.

But what if we could predict which fighter is going to win in each matchup? As matchmakers, that would allow UFC to influence each fighter's professional journey. They could, for instance, increase the odds of a fighter winning a match, while still picking an opponent strong enough to ensure audiences are entertained. UFC could also ensure that certain fighters have strong winning streaks and control a fighter's history to help increase his or her popularity and hype them for a title match. This power would give UFC an extraordinary amount of control over how their fighters are viewed by the public and would allow them to control the UFC narrative.

So that brings us to our question, is it possible to predict UFC matches? How accurately can predictions be made? And could we use this data to help us predict event proceeds?

Literary Survey

Like many sports, UFC fighting generates a lot of money. In an article written by Sports Daily [1], in 2020 a UFC fighter was able to make up to approximately $10,000,000 with an average of $160,000 for most fighters. The value of UFC has increased throughout the years, with the UFC reporting earnings of 1 billion dollars for the first quarter of 2022 [2]. Creating a predictive model for UFC matchups can optimize earnings for the UFC and encourage betting for the sport.

Many people have sought to create a valuable predictive model for UFC fighting. Although UFC fighter technique is different from one fight to another and is dependent on an opponent, one study [3] evaluated fight technique to determine the results of a head-head fight. Data such as significant strikes per minute, striking accuracy, strike defense, takedown accuracy, and more were evaluated to create a Deep Neural Network model. This model resulted in an 80% accuracy rate.

Another study [4] observed similar technical fighter data to generate a predictive model. This study observed different classifying models such as perceptron, random forests, decision tree, SGD classifier, SVM, Bayes, and KNN to see which one performed best in predicting fight outcomes. They found that the perceptron, a neural network model, performed the best out of all of them. However, this study was not able to generate a model that had anything more than 60% accuracy because the model was trained on a small, noisy dataset.

Data Collection

Data collection started with Sherdog.com. This website proved to be the most comprehensive source of UFC data and included both amateur and professional fight history as well as breakdowns of how each fight was won. For that reason, we opted to web-scrape all UFC fights and fighter data from Sherdog from when the UFC started in 1993 to the present day (2022). Fight information included when and where a fight took place, who won, and by what method. Key fighter information included each fighter's height, weight, age, locality, nationality, style, and weight class. Since Sherdog records each fight individually, we could calculate each fighter's career stats (such as win-rate and average number of rounds) at the time that each fight took place. This allows us to see each fighter's career stats as they were at the time each fight took place. In contrast, all the other approaches we discovered through our literary analysis utilized a fighter's overall career stats when predicting past fights; meaning their models "could look into the future" and see how many fights each fighter will go on to win, which is not realistic and means their models are over-informed.

We utilized a second source of fighter data scraped from ufcstats.com. This website details each fighter's offensive and defensive stats, like how many significant strikes they have landed or how many takedowns they have blocked or dodged. This was matched to the Sherdog data by either the fighter's name or nickname but was more troublesome than that. Many fighters had multiple spellings for their names, possibly for different languages. Other fighters put their last name first in one dataset but not in the other. Some even had typos. In general, we tried to match fighters by first and last name, if that didn't work, we attempted to match by nickname and validate that their first and last name were similar (since there are cases where multiple fighters have the same nickname). If that failed, then we had the computer pick best-guess matches based on string similarity on both name and nickname and matched entries by hand.

In addition to fight and fighter data, we collected data on the locality where each fighter trains and where each event took place. Locality information was pulled from NASA's Socioeconomic Data and Applications Center (SEDAC) dataset containing census information and various demographic estimates for every region for 2000, 2005, 2010, 2015, and 2020 (which could be as detailed as an entry for a single city block). This dataset includes numerous predictors, including the total population, population by sex and age, land area, water area, latitude, and longitude for each location. The population data was averaged and divided by total land area to convert these estimates to average population densities. We used this data in our models to not only determine if environmental or demographic factors are correlated with fight outcomes, but also to help explain UFC pay per view (PPV) purchases.

Attempting to connect a fighter's training location with the locality information from NASA proved difficult. Most foreign cities have local names as well as English names, meaning that the same location could be written differently in the NASA data and the Sherdog data. To address this, an additional world-cities dataset (from simplemaps.com) was used to identify the global coordinates of cities by their English name. To link the two datasets, an approximate nearest neighbor's index using Haversine distance was utilized to create a one-to-many mapping to the NASA dataset. However, even between these two sources, there were still some localities that were too obscure to find. In these cases, either the matching entries were found manually, or the average for the higher-level administration name was used. Meaning, if a city and Virginia couldn't be found, the average for the entire state was used.

After combining all datasets, there was the matter of validating that it was combined correctly. This initially involved random sampling of data entries and comparing against the source material, but some more involved methods were utilized as well. To ensure that localities were taken from only one source (not from multiple places around the earth with the same name), the range of longitudes and latitudes found for each locality was compared against the calculated range of coordinates that could be covered by a square land mass with the same land area. Each entry was allowed to have a longitudinal or latitudinal range twice that of the side-length of the calculated square. That leniency factor of 2x was chosen in order to account for the fact that most regional boundaries have a substantial eccentricity. All localities that failed this test were manually reviewed or fixed.

Data inferencing was done on a case-by-case basis. Predictors like height, weight, age, and locality information were inferenced from fighter averages while a fighter's reach was inferenced from a simple linear regression on height. To keep things simple, missing career data was inferenced from the average of every fighter at every point in their professional career and assumed to remain constant over the duration of the inferenced fighter's career.

In the end, our cleaned dataset has 6,783 fights, 2,347 fighters, 217 events with disclosed PPV data, and 163 potential predictors (after the creation of dummy variables). The response for each fight is 1 if fighter A won the match, 0 if fighter B won, and 0.5 on a draw. Which fighter gets to be fighter A is arbitrary and was purposefully shuffled to ensure a baseline accuracy of ~50%. Rows were also shuffled to prevent data from being stored chronologically. The first 80% of data was used as the training set, the next 10% was the validation set, and the remaining 10% was the test set.

Exploratory data analysis of the dataset showed that the most frequent win method was by unanimous decision, which means the judges decide who wins the match after all rounds are over and both fighters are still standing. Also, the average number of rounds is 2.3; only three rounds occur in non-championship UFC fights, so many of these fights last almost the whole allotted time. Although there are 94 different nationalities amongst the fighters, the two most common are the United States and Brazil. Around 53.3% of the fighters are from the United States and around 12.5% of fighters are from Brazil. Since so many fighters come from the same areas, we ended up lumping many of the less common nationalities together into the base-case 'other' to reduce the number of predictors. Similarly, we created dummy variables for fighting style for only the top styles and the rest in the base-case as 'other.'

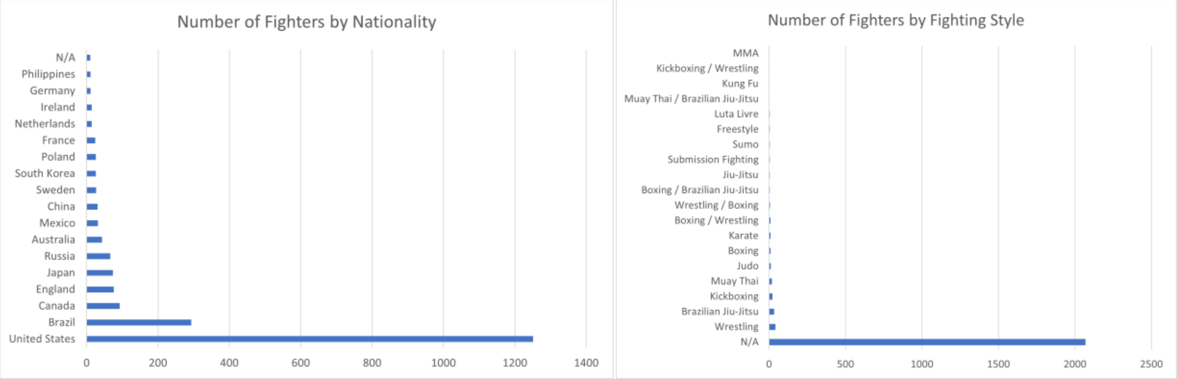

Figure 1 This shows the number of fighters by each nationality. For brevity, only nationalities with more than 10 fighters are shown. N/A is for fighter's who did not define their nationality.

Figure 2 Shows the number of fighters by their identified fighting style. Only styles with more than 3 fighters were included. Note that it is most common for fighters to not identify as having a style, probably because this is MMA and most fighters don't have just one style. Also note that some fighters identified as having MMA as their style, we kept them separate from N/A since we don't have a concrete distinction between the two, and MMA as a style is uncommon with only 3 fighters.

Pay-Per-View Data

A big portion of UFC revenue comes from pay per view (PPV). A user can pay to watch a single event or pay a monthly subscription on platforms such as Hulu or ESPN+ to watch multiple fights or live events. The current dataset has multiple fights from the same event, so an interesting question arose: can the PPV for an event be predicted based on the differences in the fighter's height, weight, locality, etc.? This resulted in another dataset, with each row containing averaged information about a past event, the resulting PPV, total attendance, etc. In order to combine this dataset with the cleaned dataset, we grouped the cleaned dataset using Python by event to obtain the overall average absolute difference in fighter A and B's height, weight, most common nationality, etc. from all fighters for that event. Finding the most common values for categorical variables, such as nationality, was challenging since there were more than 50+ nationalities in the dataset. Unlike the head-to-head data, the strategy chosen to reduce the number of nationalities was to choose the top 10 most occurring ones and set the rest as 'other'. A similar logic was applied to the other categorical variables. The combined dataset has 217 observations.

Methodology/Results

Deep Neural Network (DNN)

The architecture of our first model consists of a single DNN to predict the relative rating of both fighter's A and B, using the same weights and biases for both fighters. This method of using the same weights in parallel is referred to as a Siamese network. We then take the difference in the predicted rating for A and B and interpret that as the logit for fighter A winning. Next, we feed that predicted value through the inverse of the logit function (the sigmoid function) and directly regress on our output. This architecture has the advantage of having a meaningful intermediate result: the output of a single channel of the Siamese network is the relative rating scale of an individual fighter. This number corresponds to how good an individual fighter is and can tell us who are the best and worst fighters of all time.

Three variants of this architecture were created: one with zero hidden layers, one with a single hidden layer, and one with two hidden layers.

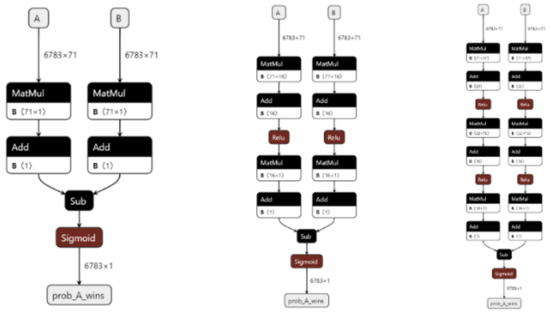

Figure 3 On the left is a network with no hidden layers. In the middle is a network with a single hidden layer. On the right is a model with 2 hidden layers.

Backwards stepwise regression was used for parameter selection. And the results of the parameter selection showed that professional wins by a given method (ko, tko, submission, etc), locality longitude and latitude, and locality information by gender caused the model to overfit to the training data and were removed. These variables were removed when training the final models.

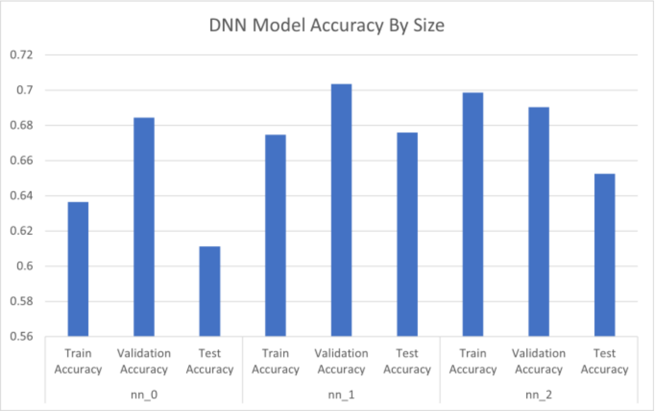

Figure 4 Accuracy was determined by the rate at which the models correctly predict a win for fighter A. Predicting draws is a later goal for threshold optimization. Note that training set accuracy increases with model depth, but validation and test results peak with 1 hidden layer, suggesting that deeper models simply over-fit to the training data.

The architecture with a single hidden layer performed best on the validation set with a 70.4% accuracy at predicting a win for fighter A. It is likely that this model, denoted as nn_1, performed best because it was able to leverage the benefits of a hidden layer without overfitting to the training data.

Lastly, we can feed individual fighter data through the Siamese channel of this model to predict who the best and worst UFC fighters are. Weirdly enough, none of the top 10 predictions were mentioned in any top 10 UFC fighters lists, like Clutch Points top 25 UFC fighters of all time [8], yet their stats were quite impressive. Some of these fighters aren't even the first result on Google when you search their names. Maybe there is an action to be taken by UFC to better look into these fighters? Links to the Sherdog sites for the full top 10 predictions can be found in the appendix.

Logistic Regression

Given that our research question involves attempting to accurately predict fight results, we felt that a logistic regression model was an appropriate choice as there is a binary response variable (fight won or lost). The model focused on predicting whether a fighter would win or lose a fight against a given opponent based on data and statistics about the fighter and their opponent. The original dataset used for this model was the "Fights_with_fighter_deltas_dataset.csv" dataset (n = 6783). Before building the model, draws were filtered out of the dataset in order to create a binary response variable (win = 1, loss = 0). Since the proportion of draws in the dataset was very small (< 1%), and we were not concerned with predicting draws at this point, we did not feel that their exclusion would have a negative impact on the model. This gave us our final dataset to be used in building the model (n = 6732).

The logistic regression model was focused on predicting results based on the comparative variables between the fighter and their opponent. So, for the first run, all comparative variables related to fighter physical characteristics/fight records/match statistics were included, as well as locality demographics by population density. The model was created using the training set, and we then analyzed the predictions on the validation set. The threshold we initially used to predict a win was p >= 0.5 and resulted in a validation set accuracy of 68.2%.

In analyzing the variables, there were some interesting observations. No variables related to a fighter's amateur fighting history had significant predictive power ($p < 0.05$). Also, age in relation to the opponent (older fighters more likely to win) and number of previous fights (less previous fights more likely to win) were both variables with very strong predictive power ($p < 0.001$), indicating that older fighters without a long fight history would tend to fare better. A VIF analysis revealed a high degree of multicollinearity amongst several amateur fight history variables. Given that none of these variables had significant predictive power, they were removed from the model for the second run.

After removing the amateur fight history, a second model was created using the training set, and the accuracy of the predictions on the validation set was calculated. Interestingly, the accuracy on this model with a reduced set of variables at this threshold was worse than the original model at only 67.3%. Given these results, Model 1 using a threshold of 0.5 was chosen as the best logistic regression model, and its accuracy was measured against the test set to obtain an estimate of real-world performance. The model performed with an accuracy of 66.5% on the test set, slightly lower than its performance on the validation set.

Random Forest

The final model created was a random forest regression model. The model type was selected because it performs well on high-dimensional data, it is a commonly used ensemble technique that is easy to explain, and it could be trained with and without feature selection. Accuracy for the random forest models was measured by the true positive rate for a win for fighter A (with a threshold of 0.5) on the training, validation, and test datasets.

The first random forest model created contained all features known before a fight as discussed in the Data Collection section. Hyperparameter tuning using a randomized grid search for 6 hyperparameters was used to identify the hyperparameters of the final model. Grid search indicated {'n_estimators': 1155, 'min_samples_split': 10, 'min_samples_leaf': 5, 'max_features': 'auto', 'max_depth': 90, 'bootstrap': True} were the best features for a random forest model with all features. The hyperparameter tuned model outperformed the non-tuned model by about 1% (65.5% accuracy for tuned versus 64.8% accuracy for non-tuned).

Since the original random forest model contained 75 features, it would be difficult to interpret the results. For that reason, a second random forest model with Scikit-Learns feature selector ("Select From Model"). This method identified 28 features that had the greatest correlation with predicting wins or losses. Several of these features included the difference in height, age, number of fights, wins, KOs, TKOs, submissions, losses, and strikes landed per minute. This simplified model was able to predict with 64.7% accuracy the result of a fight. Since this accuracy was slightly lower than the non-feature selected random forest model, the random forest model with all features was selected as the final random forest model that is compared to the other two models in the Model Comparison section.

Threshold optimization

So far, we have only been able to measure model accuracy by how often each model correctly predicts a win for fighter A. However, there are three outcomes to this problem: a win, a loss, and a draw. We would expect that for a model to predict draws, the output would be a probability close to 50%. And since the problem is inherently symmetric (meaning the model should output the same odds for a fighter winning regardless of if they are fighter A or B), the threshold around this 50% mark should also be symmetric. We will define the draw threshold as the amount above or below the 50% mark needed to predict a win or loss. So, for example, we can choose a single draw threshold of let's say 1%, which would mean that any prediction with an output higher than 51% would mean fighter A wins, an output with less than 49% would mean fighter B wins, and anything in-between would be a draw. To find the best draw threshold, we can pick values between 0 and 1 at a fine interval and test how the f1-score changes on the validation set. Specifically, we will measure the weighted f1-score since this is a three-class problem and the traditional f1-score is for binary outputs.

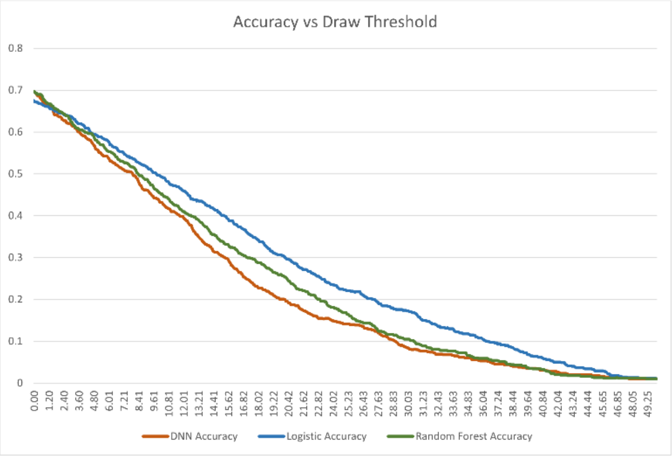

Figure 5 Weighted F1 score on the validation set only drops as we increase the draw threshold. The draw threshold is measured as a percentage.

The weighted F1 score is highest when the draw threshold is 0% for each model, meaning that we do not benefit from attempting to predict ties. This is probably because many of the decided fights end up having predictions close to 50% and distinguishing them from draws is nearly impossible. Additionally, there are very few ties in the validation set: only 7 out of 678, meaning there is little reward for getting them right. Because of this, all accuracies from here on will be measured simply as the rate at which a model correctly predicts a win for fighter A (meaning wins are measured as a prediction >50%, while ties or losses predictions less than or equal to 50%). Luckily, this is similar to how other projects from our literary survey recorded accuracy, making our results more comparable.

Model Comparison

The three models discussed above were compared based on their accuracies on the training, validation and test data sets. The results on the test dataset were ultimately what was used to select the best-performing model. The figure below highlights the training, validation, and testing errors for each of the models.

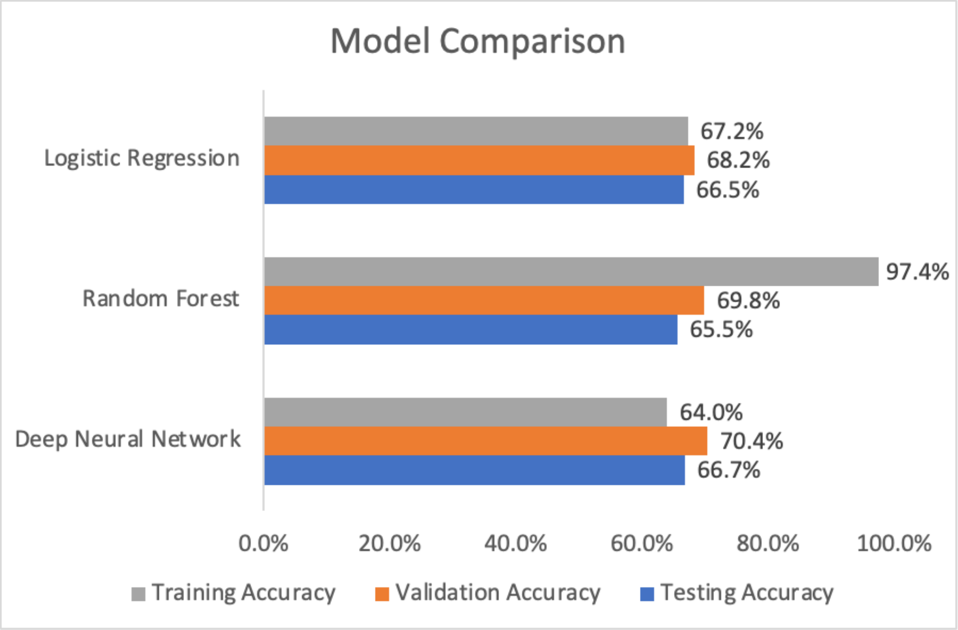

Figure 6 Comparison of each of the three models at predicting who would win a UFC fight

The deep neural network model came out on top with an accuracy of 66.7% on the test data, barely edging out logistic regression which had an accuracy of 66.5%, and random forest which had an accuracy of 65.5%.

There were several interesting takeaways from the models. Deep Neural Network, while being the most accurate, was only 0.2% more accurate than logistic regression. Deep neural networks are a black box in that their parameters can be hard to interpret. For that reason, logistic regression would be the optimal model to present to a less technical audience.

Logistic regression had several interesting takeaways including that a fighter's amateur fight history was not significant in predicting if they would win a fight. Logically you'd think if a fighter performed well as an amateur, they may be more likely to perform well as a professional, but that wasn't the case. Also, age had a strong positive correlation with wins, and the number of previous fights a fighter had had a strong negative correlation. Counterintuitively, that means older fighters and those with fewer previous fights were more likely to win.

Random forest performed better on test data after hyperparameter tuning but interestingly performed worse after feature selection, so most of the features were used on the final random forest model. This likely led to the next point in that random forest performed well on training but had a large decrease to the testing data indicating it was likely overfitted to the training data.

The models were all able to predict UFC fights with about 66% accuracy. Also, the logistic regression model showed there is a correlation between physical features and the result of a fight.

Notes Against Other Head-to-Head Attempts

As mentioned in the literary analysis, other researchers have reported higher accuracies than what we have been able to achieve. Most notably a Towards Data Science article by Filipe Cunha [3] suggested that it is possible to achieve an impressive 80% accuracy when predicting a win for fighter A. However, a deeper look into how the data was collected suggested there was a flaw. Normally, the choice of who gets to be fighter A is arbitrary. In a properly shuffled dataset, the number of entries where fighter A wins should be ~50% and should not be correlated with any individual fighters' stats but rather only correlated with the difference between two fighters' stats. However, most historical data on UFC events will always record the winner as fighter A. This means that an unshuffled dataset will result in a model that predicts an outcome of 1 for every fight regardless of the fighter's stats, which is not very useful. Looking into Filipe Cunha's GitHub code for this project, we found that he attempted to remedy this by taking the latter half of his data and swapping fighters A and B. However, what he failed to notice was that his data was organized chronologically, and each fighter's birth year was a predictor. Meaning that fight outcomes could be predicted by when just one of the fighters was born, which is not a real relationship. We created a very simple logistic regression model in R on his data with only one predictor, the birth year of fighter A, to see just how much of a false advantage this lack of shuffling provided. This model provided a P value of <2e-16 on fighter A's birth year and achieved an accuracy of 73.3% on the training data. This is indicative of false bias in his dataset and the reported 80% accuracy is likely over-inflated. We could only find one other attempt that claims to out-perform ours [7], which claims to reach a 73% accuracy on their validation set. However, they utilized a fighter's full career stats to predict past fights. Meaning their network could essentially see into the future and know the fighter's overall win rate, making it over-informed and thus impossible to draw conclusions about their model's predictive capabilities.

Pay-Per-View Models

When evaluating our pay-per-view data set, we had many variables to evaluate our PPV value. We first performed stepwise linear regression to find variables that contributed more to the model than others. After performing stepwise linear regression, we found the following variables minimized the AIC (Akaike Information Criterion) value of the model the most:

- fighter_amateur_average_number_of_rounds_a_minus_b
- strikes_blocked_or_dodged_percent_a_minus_b
- fighter_amateur_percent_wins_by_technical_submission_a_minus_b
- stance_none_a_minus_b
- fighter_professional_percent_wins_by_tko_a_minus_b
- fighter_amateur_percent_losses_by_ko_a_minus_b
- nationality_brazil_a_minus_b
- fighter_amateur_percent_losses_by_other_a_minus_b
- fighter_amateur_percent_wins_by_other_a_minus_b
- style_kickboxing_a_minus_b

We only had 217 observations for our PPV value, so to design a more accurate model we chose to use a binary variable for our dependent variable which represented whether an event generated a PPV value greater than the median PPV values (represented with a 1) or below the median (represented with a 0). This allowed us to use logistic regression and random forest to predict the binary outcome.

Since the sample size is relatively small (217 observations), we used Monte Carlo cross-validation to be more confident with the performance of the models. For each of the 100 iterations, 20% was randomly assigned to testing and the remaining 80% was allocated for training. This split was chosen since it is common practice in machine learning to perform an

80/20 split. The test error, accuracy, and variance were calculated and averaged across all iterations. Below shows the average performance metrics for both models:

	Logisitic Regression	**Random Forest**
Test Error	34.85%	27.43%
Test Error Variance	0.45%	0.37%
Test Accuracy	65.15%	72.57%
Test Accuracy Variance	0.45%	0.37%

We can see that random forest has a higher test accuracy than logistic regression. The test error was obtained by counting the number of instances the model incorrectly predicted the class and dividing it by the total number of observations. It should be noted that logistic regression outputs a probability between 0 and 1, so we set p >=0.5 as 1 and p < 0.5 as 0.

We experimented with utilizing the head-to-head fight predictions as an input to both models on PPV to get an idea if close matchups yielded higher profits. The coefficient value was large for logistic regression, and even suggested that an event with only one-sided matches would lose nearly all viewers, but we found that the values were not very significant (p-value > 0.1), so the results were dropped from the final model.

Conclusion/ Next Steps

In the end, we were able to show that UFC fights can be predicted in many cases. Meaning that our models can help UFC improve their ability to make closer matches and help crown the best UFC champion. We also were able to leverage our data to predict UFC proceeds from PPV purchases. It should be noted that our data is not perfect, since all head-to-head models exhibit varying degrees of over-fitting. It might be the case that our validation or testing sets were too small and contained random effects that threw off our results. As a result, in-depth cross validation on each of these models may prove useful moving forward. Also, it is within our best interests to try to expand our dataset as much as possible to better represent all types of matchups. We noticed that ufcstats.com has more recorded fights than Sherdog, suggesting there is a hole in Sherdog's records that we could attempt to fill. Another simple addition would be for UFC to provide more disclosed PPV data, since the most recent event in that dataset is from only 2017. All in all, although we can definitively provide useful insights for UFC today, there is still more work to be done to fully leverage these models to increase UFC's profits.

Appendix

- Top 10 predicted UFC Fighters in increasing order
  - https://www.sherdog.com/fighter/Ottman-Azaitar-173073
  - https://www.sherdog.com/fighter/Aliaskhab-Khizriev-209647
  - https://www.sherdog.com/fighter/Natalia-Cristina-da-Silva-193033
  - https://www.sherdog.com/fighter/Cristiane-Justino-14477
  - https://www.sherdog.com/fighter/Georges-St-Pierre-3500
  - https://www.sherdog.com/fighter/AbdulKerim-Edilov-63045
  - https://www.sherdog.com/fighter/Tom-Aspinall-65231
  - https://www.sherdog.com/fighter/Zha-Yi-228277
  - https://www.sherdog.com/fighter/Jeong-Yeong-Lee-135897
  - https://www.sherdog.com/fighter/Abusupiyan-Magomedov-60755

Works Cited

1. https://thesportsdaily.com/news/2021-ufc-fighter-salaries-complete-list-fox11/
2. https://www.si.com/mma/2022/06/10/report-ufc-makes-over-1-billion-per-year
3. https://towardsdatascience.com/predicting-ufc-bouts-with-dnn-classifier-f955e9abe6c6
4. https://www.kaggle.com/code/calmdownkarm/ufc-predictor-and-notes
5. https://towardsdatascience.com/predicting-ufc-fights-with-machine-learning-5d66b58e2e3a
6. https://arxiv.org/pdf/1712.03686.pdf
7. https://medium.com/@yuan_tian/predict-ufc-fights-with-deep-learning-e285652b4a6e
8. https://clutchpoints.com/ranking-the-25-greatest-ufc-fighters-of-all-time