Insights into UFC Data from Machine Learning Applications

Drew Robinson, Jason Willett, Nikhar Patel, Michael Magnuson, Taylor Walicki

It is **ok** to use an anonymized version of this report as an example of a great project for future classes.

Summary

Ultimate Fighting Championship, most commonly known as UFC, is a league that has continued to grow in popularity even during Covid conditions so insights into how fighters compare to one another, the ability to predict the betting odds for a given fighter, and the ability to predict the method of victory for a fight could prove very valuable. For our results, k-means clustering techniques showed a split of fighters where one group had much higher averages on a fight-by-fight basis for strikes, takedowns, and many other metrics indicating these were more aggressive fighters or their numbers were somewhat inflated by having matches that lasted more rounds than a typical fight. Community detection algorithms revealed that most male fighters were directly or indirectly linked to one another in our dataset through direct matchups or mutual opponents. For clarity, “indirectly linked” here means that two fighters didn’t directly fight each other, but they may have fought the same third fighter to form a connection or the link may come from a line of even more connections. Our linear regression techniques to predict the betting odds for “red” corner fighters yielded a mean squared predictive error (MSPE) that was nearly a quarter of the MSPE for predicting the odds for “blue” corner fighters. This makes sense since red fighters are usually the clear favorites in their fight. Finally, classification techniques for the method of victory (not including who won) proved reasonably successful with our LDA model predicting with around 75% classification accuracy showing that round-by-round data did have some predictive value for projecting the method of victory.

Motivation

Our motivation to focus on the UFC stems from a mix of practical and personal factors. On the practical side, UFC has a massive social media following (surpassing the NFL on Instagram) and is continuing to grow in terms of league revenue and viewership. Anything that is drawing this kind of attention and the corresponding revenue is worth investigating. Specifically, the gambling scene is especially vibrant in UFC and, while we don’t advocate for gambling, we hope to better inform people’s decisions if they choose to do so. We also found minimal research combining machine learning techniques with UFC data which we saw as an opportunity to do something new and innovative. Regarding our personal motivations, UFC is intriguing to analyze due to the lack of equipment used. Fighters aren’t even required to wear specific gloves like in boxing. UFC also stands apart from many other sports since the outcome of each fight is determined almost entirely by the fighters themselves with some input from the referee, and, when necessary, the judges’ scoring. All-in-all, UFC has less external factors or confounding variables compared to other sports so it seemed well suited for analysis.

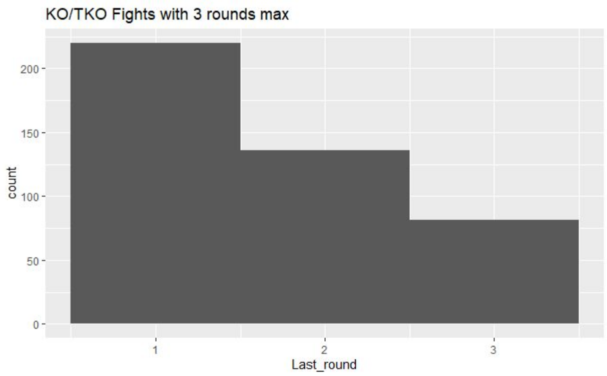
After considering all of this, we knew that machine learning may give us some insights into better understanding UFC. Unsupervised learning techniques could help us compare different fighters and better understand how we could categorize fighters using a quantitative approach. In addition, supervised learning methods could bring additional value by trying to predict metrics and class labels of interest including the betting odds for a fighter and what the method of victory was based on career data for each competitor.

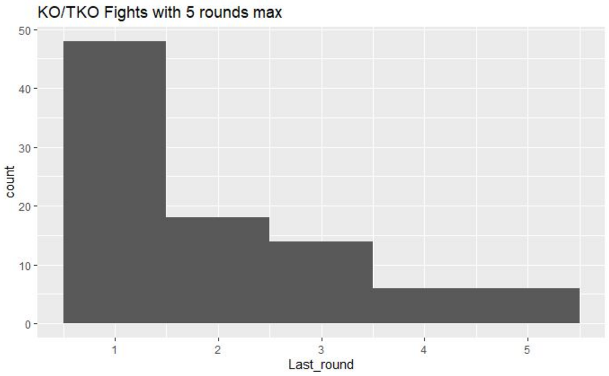
Data

We found and merged multiple UFC datasets to fulfill the questions we came up with through our own knowledge of UFC and exploration of the data we started with. Our inspiration for the project began with a UFC dataset that gave minimal location and timing information as well as career averages for the “red” corner fighter and the “blue” corner fighter going into a new bout. The dataset also listed which fighter ended up winning the match, but did not include how the fight was won whether by TKO/KO, submission, or decision. For clarification, the “red” corner fighter typically refers to the champion or the odds-on favorite. The colors are assigned to make information around the fight, particularly for betting purposes, more clear. This dataset required some cleaning and gave us sufficient data to build a fighter dataset by splitting each row (representing a fight) into new rows which each stood for a single competitor and their career averages. We took the most up-to-date career averages for each fighter and dropped the rest. However, we were also interested in finding the betting odds for the red and blue fighter before a fight which we found in another dataset. We merged this with the first dataset by date and the name of the red fighter. This process was complicated by one of the datasets not storing the dates in a user friendly format with a standardized character length, but we solved this issue using an algorithm to reformat the dates. Finally, we found one more dataset that gave round-by-round data for each fighter’s previous fight going into their current matchup. We chose this dataset since it gave a specific method of victory for the fight e.g. TKO/KO, submission, or decision. This new data ended up reshaping our interests after some exploratory data analysis. We ended up not caring as much about whether red or blue won, we wanted to know how they won for our predictive modeling goals.

Exploratory Data Analysis

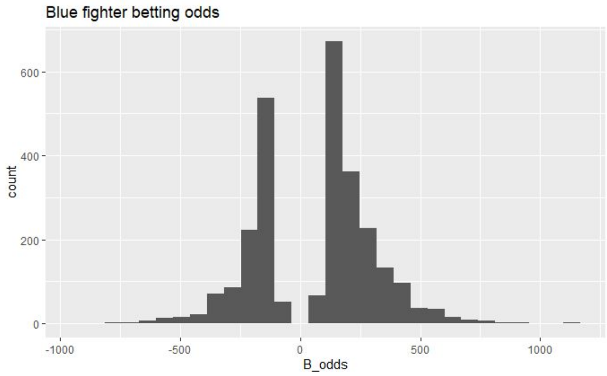
To search for further inspiration for our project, we tried multiple visualizations across potential response variables in the round-by-round data to see if these variables had a relationship with any predictors. The first graphs that stood out were two histograms with each taking a count of how many fights ended in a knockout (KO) or a technical knockout (TKO) for each round. A TKO is where a fighter is still conscious, but the official ends the match since they deemed that the match could not be continued safely for the losing fighter. Our visualizations of the methods of victory required two separate histograms since most of these fights were limited to a maximum of 3 rounds, whereas the rest could go up to 5 rounds.

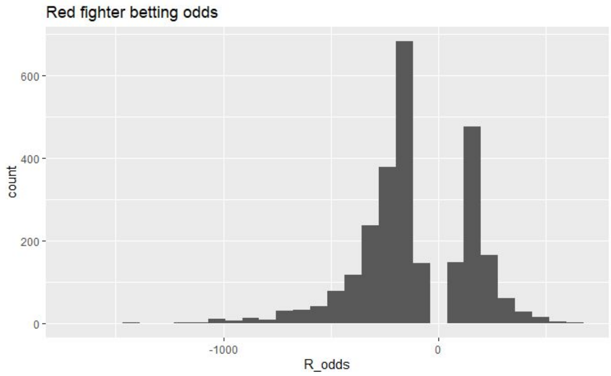




Both showed a right-skewed distribution where fights that ended in a KO/TKO typically ended earlier rather than later. This inspired us to see if an analysis of round-by-round data might give us the ability to predict the specific outcome of a fight such as the KO/TKO results shown above. Additionally, we hoped this analysis may give insight into why KO/TKO results frequently occur in the early rounds of a fight.

We also examined the distribution of the betting odds for each fighter to see if there were any trends that stood out. Betting odds with a negative value reflect the favorite, while betting odds with a positive value represent an underdog. The betting interpretation of a negative value, x, is “you must pay $x to win an additional $100 if this fighter wins.” On the other hand, the betting interpretation of a positive value, y, is “you must pay $100 to win an additional $y if this fighter wins.”





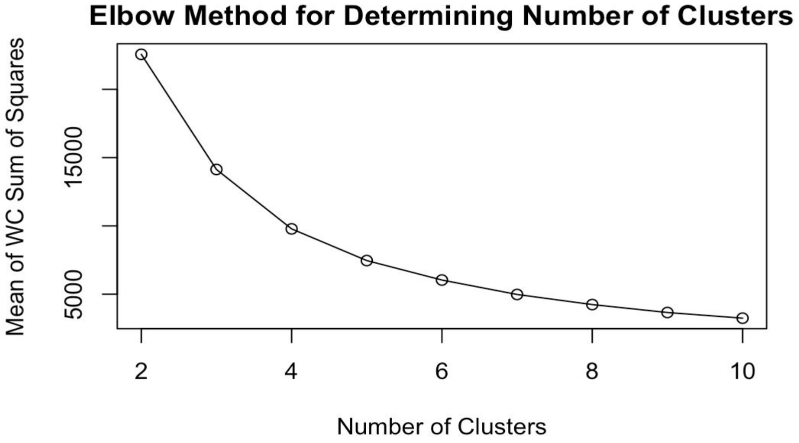
The betting odds for the red fighter support our earlier assertion that, in general, the red fighter is favored to win a given fight. This begged the question, “Is this skew in betting odds based on actual fighters’ data or are there other external factors at work here?”

Unsupervised Clustering of Fighters

Grouping fighters may go a long way in helping determine fight outcomes, viewership, fighter success, and many other factors. Therefore, our first goal was to use unsupervised clustering techniques to help us in our exploration. In hopes of creating the most meaningful clusters, we used two clustering methods: hierarchical and k-means.

Before using our clustering techniques, it was important to set up our data matrix so that we could interpret our results better. Each row of the matrix represents one of the 1099 unique fighters. The columns included three types of variables that describe the fighter as whole. The first type of variables are the fighter’s own fight averages, like how many strikes they land on average in a fight. The next type of variables are ones that indicate the averages of the fighter's opponent against them. For example, the average amount of takedowns the fighter gives up to their opponent. The final type of variables included were those that quantify the physical attributes of the fighter, like height and reach. Clustering on these three different types of variables will allow us to look into the clusters more profoundly and to develop better interpretations.

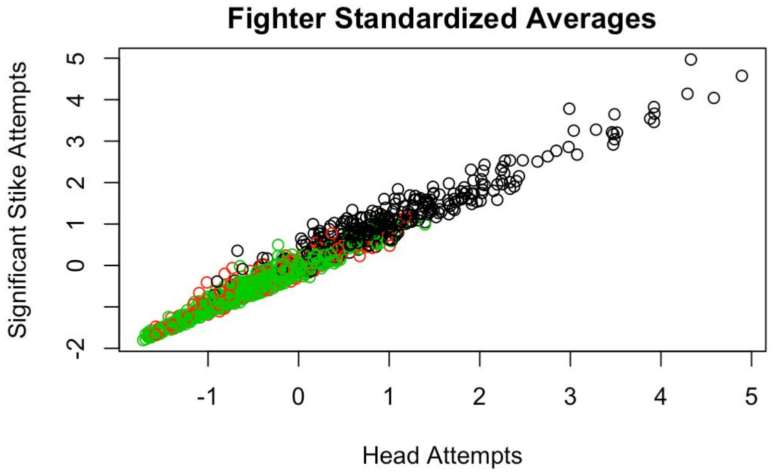
After ensuring all features had the same weight by standardizing them, it was a necessity to determine the optimal amount of clusters in order to get the best understanding. First, we plotted the mean of the Within Sum of Square values for k-mean models that had two to ten clusters:

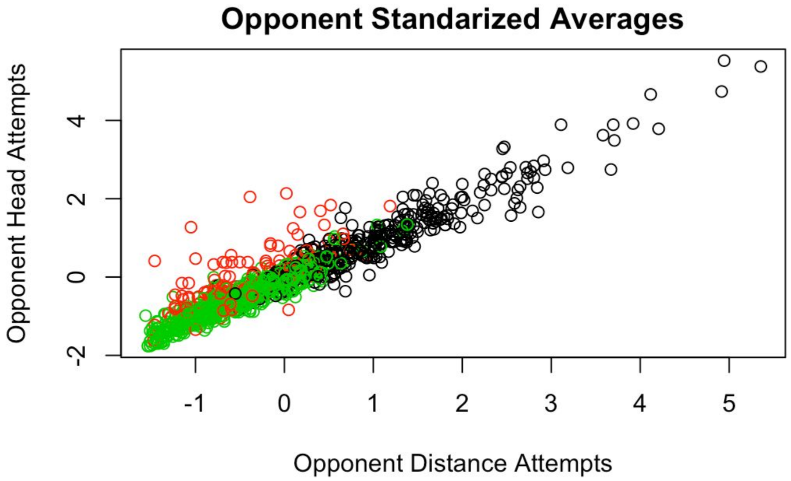


Using the elbow method, the "best" number of clusters would be at either three or four according to the graphic above. After using the elbow method, we thought it would be best to explore the silhouette coefficients for further guidance in picking an ideal number of clusters, especially since the elbow method was not entirely clear. According to the silhouette coefficients, the "best" number of clusters to use for our features is two. Therefore, we took both methods into consideration and determined that three clusters would provide us with the best findings.

Knowing our ideal number of clusters was three, we used hierarchical and k-means clustering techniques to split the fighters into this amount of clusters. After performing both, k-means clustering was by far the more interpretable method for our fighters. Hierarchical clustering led to one massive cluster and two minute groups. From k-means, we had three clusters of size 353, 153, and 593. After looking at several plots and looking at the standardized means of all the features for their respective clusters, we started to get a sense of the three different groups. We will show these plots below after explaining a few findings.

It was easiest to break down three groups by looking at the three types of variables individually. Therefore, we began by looking at how individuals from their respective groups fought on average. Evidently, there was a trend among many of these features within this type of variable. Cluster one produced higher averages compared to the other two for the majority of the features and cluster three produced the lowest averages for many of these features. For example, cluster one, cluster two, and cluster three had a mean standardized "average distance attempted" of 1.04, -0.35, and -0.53 respectively. This trend held true for the second type of variables as well, meaning cluster one fighters faced opponents who swung a lot on them and were successful. However, for the physical attribute variables, the relationship was inverse, where cluster one fighters were usually smaller and fighters in cluster three were on average bigger. Though, the means for the three clusters were all pretty close to zero, which would indicate there is slight variation in the cluster's mean fighter size. Below are some plots to help visualize the different clusters.





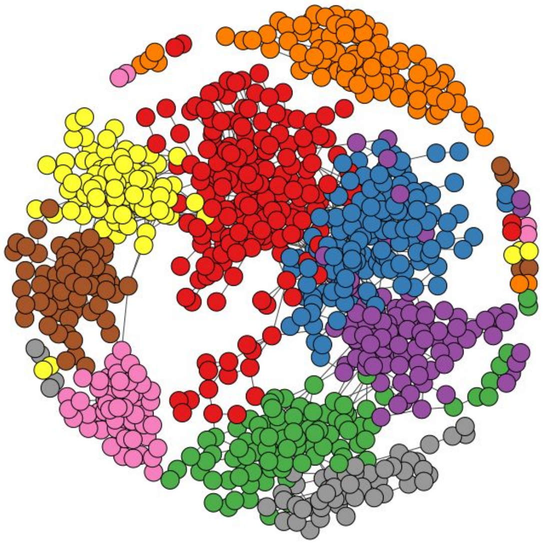
The plots above indicate two takeaways from the data. First, the relationship among the two groupings of variables are positive. Looking at the first graph, it shows that as more headshots are attempted, then more significant strikes are taken too. As for the second graph, we see that fighters who allow their fighters to attack from a further distance tend to also give up more attempts to their head. The second takeaway from these graphs is the portrayal of where these groups stand. Cluster three (green) had lower averages. Cluster one (black) had higher averages. Meanwhile, cluster two (red) had fighters mixed in the middle of it all.

Why are the averages of cluster one higher than the other two? There are two realistic explanations. The fighters either are in longer fights on average or they can be characterized as aggressive fighters. Both are valid arguments as to why their averages are higher and this is definitely an area that should be further examined. There are an endless amount of questions we can analyze with these clusters. Does one cluster create more revenue for UFC because of their fighting style or length of fight? Is one cluster more successful than the other two? With some surface level analysis, cluster one had a win proportion of .60, cluster two had a win proportion of .38, and cluster three had a win proportion of .58. These are just a few topics that should be further explored now that we know there is a clear divide between three types of fighters.

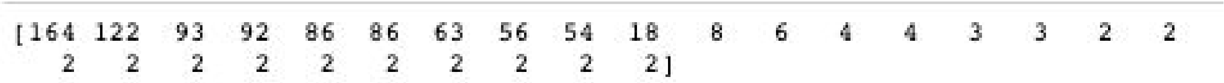
Community Detection Algorithms

Next, we wanted to bring another perspective on how to group fighters so we employed community detection algorithms. Community detection has been an increasingly popular field of research in network science due to its applications in a variety of disciplines including social, biological, citation networks. In our case, we hope to examine how communities form over a set of UFC fighters where people are linked by past fights. In doing so, we hope to be able to analyze the created communities and comment on analysis of what they could represent.
The data did not require a heavy amount of processing due to the simplicity of its inputs. The main work done was to give every fighter a unique index that would serve as replacement for creations of vertices in the graph. Edges were placed between two vertices if the two fighters had competed against one another; one should point out that due to the graph being undirected and unweighted multiple fights between the same two people were not taken into account. This means there is no weight or preference given to fighters who fought multiple times over the period.

For all the community detection done we used CHAMP (Convex Hull of Admissible Modularity Partitions) created by William Weir in the Mucha Research Group in conjunction with Louvain and the louvain_igraph package created by Vincent Traag. Both CHAMP and Louvian are heavily rooted in the optimization of modularity which is defined as, the strength of division of a network into communities where high modularity infers dense connections within a community but rather sparse connections between nodes, in our case fighters, of different communities. After processing the data, we ran CHAMP on the entire set of UFC fight data from 1993-2019 and got the below graph.



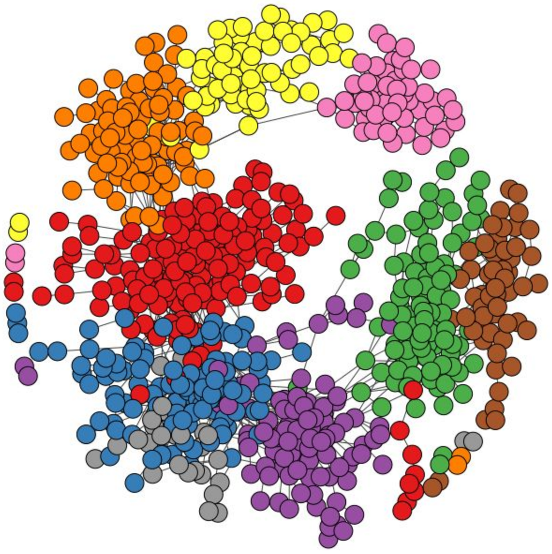
We initially ran CHAMP on the entire dataset with no filtering in terms of fights, getting the figure seen above. The created graph has a total of 886 unique nodes with 1612 total edges, a very sparse graph as one would expect in this case. In the process of creating the graph we “lost” seven edges due to having seven fights occur between the same people. One might first examine the communities created and we see that there are a total of 28 created communities (Python’s color palette does not allow for a large assortment of colors therefore there will be multiple communities shaded with the same color). Below we see the distribution of the number of fighters in each community.



```
[164 122  93  92  86  86  63  56  54  18   8   6   4   4   3   3   2   2
   2   2   2   2   2   2   2   2   2   2]
```

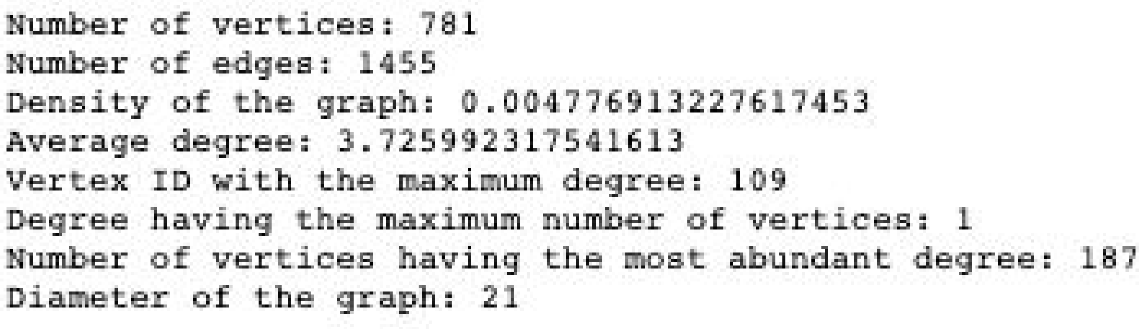
There are a few unique things one can see in both the graph and the distribution of communities. We see that there are quite a few communities of size less than 10 under; for example, if we take the community containing for “'Arjan Bhullar' and 'Adam Wieczorek' we see that these fighters only have a recorded fight against one another creating a disconnected subgraph. This pattern continues for all the communities that contain less than 10 fighters. Another somewhat more alarming fact is the large orange community containing 86 fighters seen in Figure 1 is completely disconnected from the rest of the large graph. Under observation, it became clear that the community contained only female fighters which was caused by the data set containing both male and female fighters, but no fights occurred between the two genders. Due to this, we then decided to examine the community structure of only the male fighters hoping to see interesting structure.

After realizing that in the initial graph we unknowingly created unconnected clusters, we thought rerunning CHAMP on the two genders separately would be of interest since we would expect the graph to be much more connected. In turn, we hoped this would provide much more valuable information. Below is the network of only males.



After analyzing the different communities we came to the conclusion that most were created due to the weight classes that each male fought in. This makes sense given the weight class requirements for both fighters going into a fight. However, these results from community detection algorithms contrast with the results from k-means where weight and height had a minimal effect on the clusters while the values of the career average statistics for each fighter drew clear boundaries between certain fighters.

In our new clusters, instances of smaller communities defined as containing less than 10 nodes we saw that these fighters primarily only had one to two fights in the UFC before leaving for unknown reasons. All of the communities of size two were instances where fighters had a debut fight against one another but had no further bouts in UFC competition. Below we can see some of the basic statistics of the graph below:

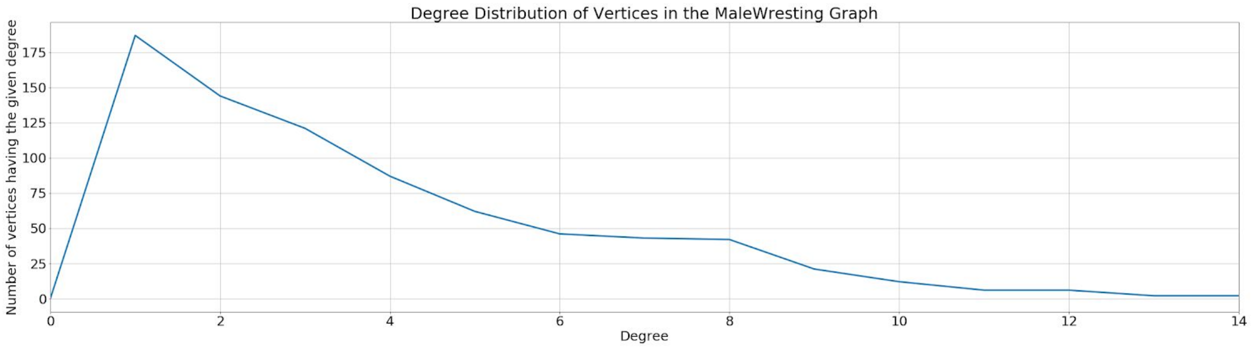


```
Number of vertices: 781
Number of edges: 1455
Density of the graph: 0.004776913227617453
Average degree: 3.725992317541613
Vertex ID with the maximum degree: 109
Degree having the maximum number of vertices: 1
Number of vertices having the most abundant degree: 187
Diameter of the graph: 21
```

We see that this graph is also extremely sparse as we would somewhat expect for this data with 781 vertices and 1455 edges. If we examine the density which is calculated as

$$D = \frac{|E|}{\binom{|V|}{2}} = \frac{2|E|}{|V|(|V|-1)}$$

We get a value of ~.0048 which is higher than the density of the total graph which is around ~.00411; although, the difference isn't extremely large it reassures us that our model is correct. A few other interesting points we can see in the data is that the fighter Ovince Saint Preux had at least 14 fights occur over this time period. Additionally, the average degree is around 3.7 which indicates that the average number of fights a male fighter had over this time period is greater than or equal to 3.73. Finally, we see that the most common degree was one with a total of 187 nodes having a degree of one. Below we can see a plot of the degree distribution which gives a generally good indication of the sparsity of the graph.

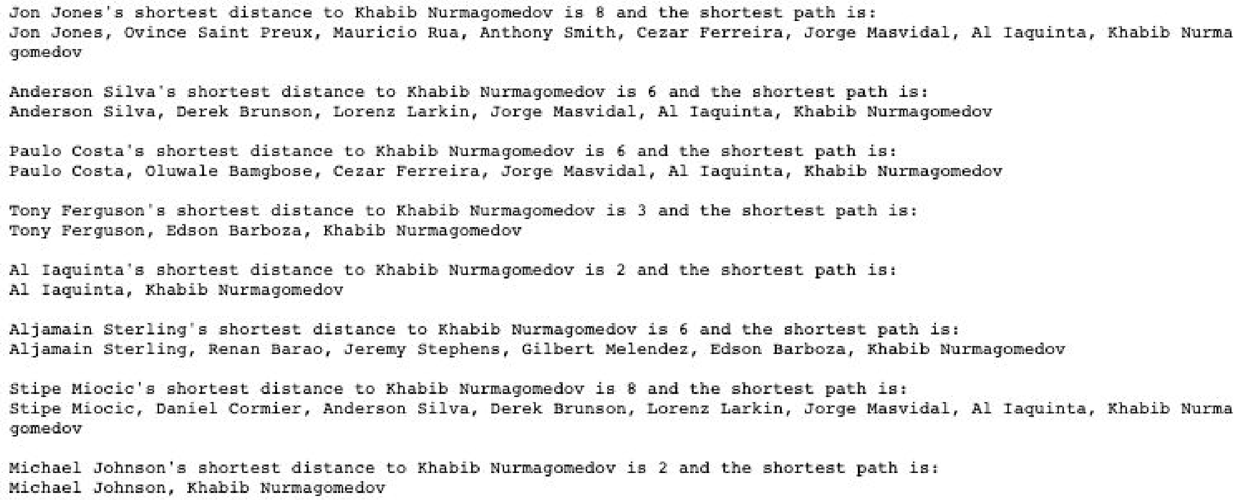


We were then particularly interested in the largest community of size 157 denoted by the large red cluster in Figure 2 and decided to rerun Louvain on just this subset of fighters which mainly consisted of lightweight fighters to see if we could if there was any underlying structure within the data. We ended up finding eight communities within the smaller subset of data which strikingly had a much higher density of .02 than both the entire data set and the male fighters. However, this is somewhat expected as we are selecting fighters that already seem to have some connection. Unfortunately, in our analysis of the new communities, we were not able to find anything super definitive besides a few instances of fighters having the same fighting style.

Some of the most interesting articulation points that we found were BJ Penn and Jim Miller. BJ Penn is rather interesting in our case as in this dataset we see that he fought in six separate weight classes and furthermore was able to win a title in two of them. Finding BJ Penn within the set of articulation points was somewhat expected under our assumption of what the communities mean due to his career fights taking place in many different weight classes. Jim Miller’s analysis was a bit more interesting, as a fighter he has spent nearly all of his fights in the lightweight division of UFC. Under further inspection into the connected nodes, we see that there exists only one fight that links him to another community which occurred through Daniel Hernandez. Hernandez only had one other fight in his career and both of his fights occurred in the Welterweight division.

A specific articulation point we chose to examine is inspired by Kevin Bacon’s “Six Degrees of Kevin Bacon.” We wanted to try and examine how many “fights” two fighters are apart and how one could create a path of connected fights to the other fighter. In this case we will look at a similar adaptation with Khabib Nurmagomedov, the recently retired, undefeated, pound for pound champion. Outside of his perfect record and impressive accomplishments Khabib’s fighting background is rather interesting as he has seldom fought outside the Lightweight division, thus connections for fighters outside of the lightweight division must occur through bridge vertices. Implementation of this was done by running Dijkstra's algorithm which is rather easy to implement for an unweighted graph. In itself, the algorithm is not of large interest but examining the “distance” of fighters could potentially allow for interesting results.

Picking which fighters to examine was largely up to our discretion so we thought it would be most interesting to see the distance of fighters at the top of other weight classes along with pound for pound champions. In the below figure you can see the output of running Dijkstras on the graph with players of our choice and target node of Khabib Nurmagomedov.

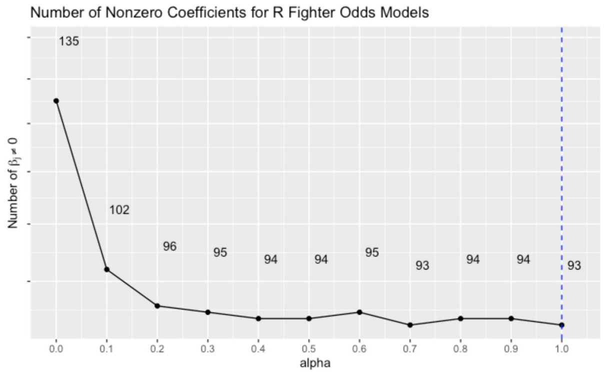


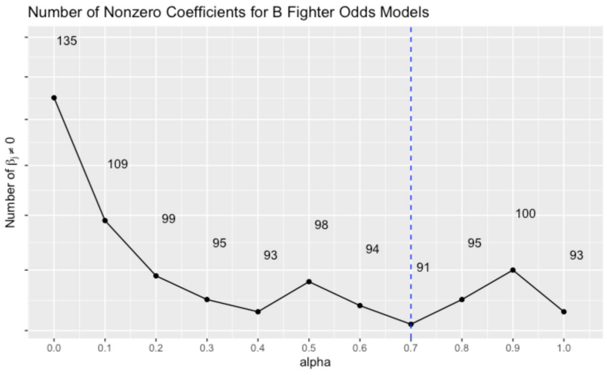
```
Jon Jones's shortest distance to Khabib Nurmagomedov is 8 and the shortest path is:
Jon Jones, Ovince Saint Preux, Mauricio Rua, Anthony Smith, Cezar Ferreira, Jorge Masvidal, Al Iaquinta, Khabib Nurma
gomedov

Anderson Silva's shortest distance to Khabib Nurmagomedov is 6 and the shortest path is:
Anderson Silva, Derek Brunson, Lorenz Larkin, Jorge Masvidal, Al Iaquinta, Khabib Nurmagomedov

Paulo Costa's shortest distance to Khabib Nurmagomedov is 6 and the shortest path is:
Paulo Costa, Oluwale Bamgbose, Cezar Ferreira, Jorge Masvidal, Al Iaquinta, Khabib Nurmagomedov

Tony Ferguson's shortest distance to Khabib Nurmagomedov is 3 and the shortest path is:
Tony Ferguson, Edson Barboza, Khabib Nurmagomedov

Al Iaquinta's shortest distance to Khabib Nurmagomedov is 2 and the shortest path is:
Al Iaquinta, Khabib Nurmagomedov

Aljamain Sterling's shortest distance to Khabib Nurmagomedov is 6 and the shortest path is:
Aljamain Sterling, Renan Barao, Jeremy Stephens, Gilbert Melendez, Edson Barboza, Khabib Nurmagomedov

Stipe Miocic's shortest distance to Khabib Nurmagomedov is 8 and the shortest path is:
Stipe Miocic, Daniel Cormier, Anderson Silva, Derek Brunson, Lorenz Larkin, Jorge Masvidal, Al Iaquinta, Khabib Nurma
gomedov

Michael Johnson's shortest distance to Khabib Nurmagomedov is 2 and the shortest path is:
Michael Johnson, Khabib Nurmagomedov
```

Supervised Regression for Betting Odds

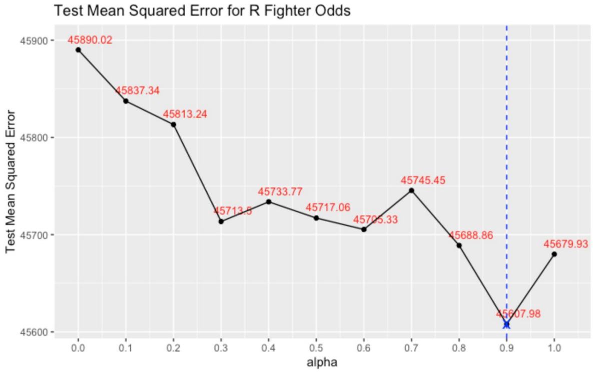
Now that we better understood the fighters themselves, we wanted to see if we could actually predict response variables for a fight given information on both fighters. We started by looking where the money was. It goes without saying that UFC fights are high-stakes. This is because of the glory associated with winning in the octagon, the prize money, and most notably, the bets that were placed on the fight. Considering the movement to legalize sports betting across the nation, it would be useful to make a model that could predict the betting odds on fighters. Sports betting is legal in 19 states, such as Nevada and New York. North Carolina recently passed a bill to legalize sports betting, but this is still awaiting approval. We decided to pursue making a model based on our UFC data that could predict betting odds for fights. The odds for the red fighter were used as the dependent variable in one model and the odds for the blue fighter was the dependent variable in the other model. This is because the red fighter is usually the favored or higher ranked fighter. The differences in the odds of the fighters are different enough to warrant two separate models.

Our approach was the same for modeling the odds of each fighter. Our dataset was not highly dimensional, but it had 149 predictors. Stepwise regression was not feasible because of this high number of predictors. Stepwise regression progressively subsets the predictors using a greedy algorithm. This means that the best single predictor of betting odds would be selected, then the next predictor added would be the one that makes the best model with the first predictor. Therefore, the final model would likely not be the best because it does not consider all possible combinations of predictors for the model. So, we decided to do the elastic net approach. This approach allows us to see which shrinkage method has the best predictive accuracy. For different values of alpha ranging from 0 to 1, we can see how effective ridge regression (alpha=0), lasso (alpha=1), and intermediate models are in terms of error. Notably, before we made the models, we standardized the variables because each variable is on a different scale. We did not want to run the risk of one variable having a larger effect because it is on a different scale. We also implemented cross validation methodology for each of these models. Cross validation is the process of creating “training” and “test” groups from your data and optimizing predictive accuracy. A model is made based on the “training” data which is usually about 75-80% of the data. We decided to use 10-fold cross validation since it is a common value for the number of folds. Then, predictions are made for the test data observations. We compare the predictions to the actual observations and compute the mean squared error. Although we chose a shrinkage method, we were interested in the number of non-zero coefficients that were present in each model.

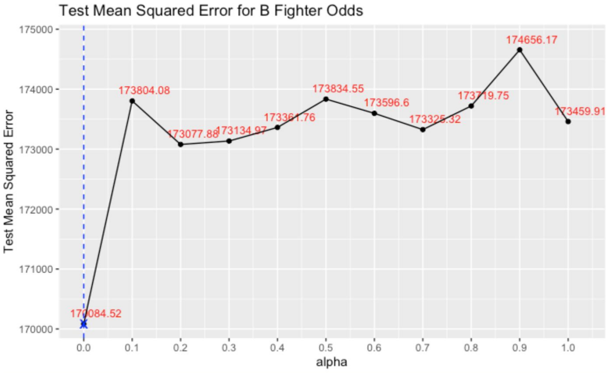

Number of Nonzero Coefficients for R Fighter Odds Models


Number of Nonzero Coefficients for B Fighter Odds Models

As shown in the figures, the model that produces the fewest number of nonzero coefficients for the red fighter is the lasso model (alpha=1) and the elastic net model with alpha=0.7. For the blue fighter, the elastic net model with alpha=0.7 produces the fewest number of nonzero coefficients. However, finding the best model is not based solely on the number of nonzero coefficients. To further decide which value of alpha would be best, we computed the test error. More specifically, we computed the mean squared prediction error (MSPE) for our predictions. The lower the test error, the better the model should be able to predict.



Test Mean Squared Error for R Fighter Odds



Test Mean Squared Error for B Fighter Odds

As shown in the figures above, for the red fighter, the value of alpha that produces the lowest test error is 0.9. Notably, its test error was only roughly 72 points lower than the lasso model (alpha=1). Combining this information with the fact that the lasso model also produced the fewest number of nonzero coefficients, we decided that the lasso model was the best model for predicting the red fighter's odds. This shows that a small subset of predictors are needed to predict the betting odds of the red fighter. In other words, a large portion of the predictors are not needed to predict this outcome. The largest predictor of the betting odds for the red fighter was the red fighter's significant striking accuracy. Significant striking accuracy is the number of significant strikes landed divided by attempts. Its coefficient value was -141.75. This means that for every increase in one significant strike percentage point for the red fighter, the odds of the red fighter is expected to decrease by 141.75 points. This finding is intuitive because as the significant strike percentage increased, the red fighter became more favored. Another important predictor was the significant striking accuracy of the blue fighter (red fighter's opponent). The value of this coefficient was 95.25. This means that for every increase in one significant striking percentage point for the blue fighter, the red fighter's betting odds are expected to increase by 95.25 points. This makes sense because as the blue fighter improves, the red fighter becomes less favorable to win the fight. One interesting finding was that the intercept was very largely negative. The value of the intercept is -161.42. This means that when all of the predictors are valued at zero, the expected betting odds for the red fighter are -161.42 which are highly favorable odds. This reflects the notion that the red fighter is considered the better fighter.

On the other hand, the ridge regression model (alpha=0) produced the smallest test error for the blue fighter. The test error for the blue fighter was substantially larger than the test error for the red fighter. This is likely because there is more variability in the blue fighter. The blue fighter is oftentimes the lower ranked fighter that is not favored. This causes more variability in the predictors based on the fighter's measurements and fighting habits. The ridge regression model has the largest number of nonzero coefficients, but this is because of how ridge regression is run. It does not delete predictors in models but adjusts their weights in the model. This is unlike lasso where predictors shrink toward zero. So, we decided that the ridge regression was the best model for predicting the odds of the blue fighter. Although ridge regression has the disadvantage of including each predictor, it is the best model in this case. This shows that a large number of predictors are important to predict the odds of the blue fighter. For instance, the most significant predictor of the blue fighter's betting odds was the red fighter's significant striking accuracy. Its coefficient value was 103.00, meaning that for each increase in one significant striking accuracy percentage point of the red fighter, the betting odds for the blue fighter is expected to increase by 103.00. Similarly, the second most important predictor was the blue fighter's significant striking accuracy. Its coefficient value was -92.54. This means that for each increase in one significant striking accuracy percentage point of the blue fighter, the betting odds for the blue fighter is expected to decrease by 103.00. These results show that the fighter with the higher significant striking accuracy is more favorable. Once again, it should be noted that there is high variability among the blue fighters and more predictors are needed to explain this variance.

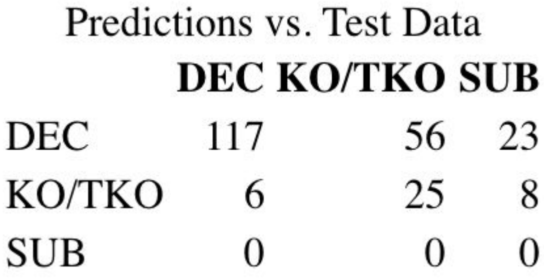
The takeaway message from this analysis is that betting odds are difficult to predict and depend on the fighter. There is more variability among the blue fighters which led to higher test errors. Red fighters are higher ranked and are usually the better fighter so predicting the odds that they will win is slightly more straightforward. The significant striking percentage of both fighters were shown to be the most important predictors of the red fighter's betting odds. However, betting odds are impacted by factors outside of the ring. This means that betting odds are influenced by factors that are unrelated to the fighters and their characteristics. Future models could be improved by touching on some of these factors and learning more about the betting odds process.

Supervised Classification of Fight Outcomes

Finally, we wanted to see if we could predict the method of victory for a fight. First, we need some background information. In mixed martial arts, there are three types of outcomes that are possible for any fight (judge's decision, knockout/technical-knockout (KO/TKO), submission). These three outcomes typically stem from differing styles of fighting, and a fight's entertainment quality can vary based on the outcome of a fight. Typically, fans find a fight that ends in a knockout very entertaining because a knockout is usually preceded by lots of action beforehand. On the other hand, judge's decisions are often considered as "boring" by many UFC fans because there was no decisive victor, and the winner of the fight will be determined by subjective judges. Due to the fact that UFC fans typically have to pay a pay-per-view fee before watching UFC fights on TV, many fans may want to know whether the $65-$85 fee is worthy of an investment. Additionally, UFC is a sport that attracts a lot of gambling attention, and the manner in which a fight is won or lost is one of the many categories fans can gamble on for UFC fights (this type of bet does not factor in who wins). Considering these two factors, we thought it would be valuable to predict the method of victory UFC fights.

The variable that we are looking at in this problem is by nature a categorical variable; there are only three possible values. Therefore, we applied a variety of supervised classification techniques on this data in order to predict the type of outcome of UFC fights. First, we standardized the data due to the fact that many of the variables were measured in different ways (some variables were continuous, e.g. body measurements, others were count variables, e.g. number of rounds fought). After that, we looked at the proportions of each type of outcome in the data set and found it consisted of: 50.1% decisions, 32.94% KO/TKO, and 16.95% submissions. We did this because we wanted to see if outcomes were fairly evenly distributed or if they were skewed one way or the other. We wanted to see if the outcomes with more representation in the data were predicted more accurately. We then split the data into a 75-25 percent train and test data set split, and then we applied k-nearest neighbors to data.

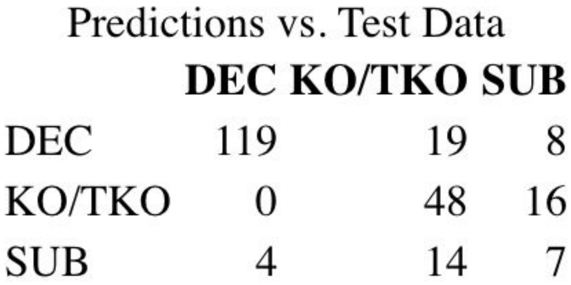
We used 10-fold cross validation on the training data set across a set of k values ranging from 1 to 499. After completing this process, we calculated and entered the mean cross validation error against the 499 potential k values into a data frame. The minimum of the errors showed that when k was equal to 67, the model produced its lowest cross validation error of 0.4450704. We then refit the model on all of the training data, using 67 as our value for k. After evaluating this model on the test data, we came up with a test error 0.3957447, meaning the model properly classified new data with about 60% accuracy. Additionally, we built a confusion matrix (shown below) to show a more in-depth breakdown of the predictions and their accuracy. The model correctly predicted the judge's decisions 95.12% of the time, KO/TKO's 30.86% of the time, and submissions 0% of the time. This model was clearly impacted by majority voting, which greatly hurt its predictive accuracy on the less represented types of outcomes in the data.



Predictions vs. Test Data

	DEC	KO/TKO	SUB
DEC	117	56	23
KO/TKO	6	25	8
SUB	0	0	0

We wanted to see if there were any other classification techniques that could improve this classification accuracy, so we next attempted linear discriminant analysis. We then fit the LDA model on the training data, using the lda function in R, and made predictions based on the test data. The resulting confusion matrix (shown below) shows the accuracy of the LDA model for the 3 different types of fight outcomes, which had an overall classification accuracy of about 74.04% (a noticeable improvement over k-nn). Additionally, this model improved on predicting all types of outcomes. The judge's decisions were predicted accurately 96.75% of the time, KO/TKO's were predicted correctly 59.26% of the time, and submissions we predicted correctly 22.58% of the time. This model still showed a bias towards judge's decisions, but it performed much better on the smaller categories than the K-nn model did.



Predictions vs. Test Data

	DEC	KO/TKO	SUB
DEC	119	19	8
KO/TKO	0	48	16
SUB	4	14	7

We wanted to be thorough in our investigation to predict types of fight outcomes, so we also attempted a QDA model. When trying to run our QDA model with the training data, we encountered an error. After doing research into the potential cause of this error, we determined that the multiple categorical variables in the data were causing this error, so removed those variables and again attempted to run the model. Unfortunately, we began to run into a different error that said, “rank deficiency.” We spent some time researching the meaning of this online, and one popular explanation that we saw is that the problem could be with the entire data set itself (either there isn’t enough data or there is too much of one type). Another potential explanation was that there was too much collinearity between certain variables in the data set. We tried to find certain variables with high covariances to fix this issue, but, due to the high number of variables in our data (approximately 850), we weren’t able to fix this potential problem. Disappointingly, we weren’t able to find a solution to the QDA error within the timeframe of this project, so we had to settle for k-nearest neighbors and LDA for our techniques.

We concluded that we were able to successfully classify the type of outcome of a UFC fight with some decent accuracy. Our LDA model was able to correctly predict the outcome of almost 3 out of 4 fights, so we certainly think that this model would hold some value when trying to gamble on the outcome of a fight. The success of this model may also reflect certain similarities across the class labels regarding the covariance matrix since LDA assumes that the covariance matrix is the same for all of the class labels. Additionally, this model could give a consumer a pretty good idea of whether or not a fight may be worth watching and thus paying money for depending on what result they are hoping for. Unfortunately, we weren’t able to overcome the models’ bias toward the judge’s decisions, which hurt the overall accuracy of our models. In the future, we would like to find either more robust data or better classification techniques for this data that could help further improve the accuracy in our predictions. Having that said, we are happy with our LDA model, and believe it could be useful to both UFC fans and gamblers.

Conclusion

In summary, we found two very different results from our clustering techniques. K-means clustering emphasized the difference in the means of many different statistics averaged on a fight-by-fight basis compared across all the fighters in our data. Alternatively, community detection algorithms split fighters mostly by weight class, with a few articulation points connecting all of the male fighters. We also found different regression models for predicting red and blue fighter betting odds based on cross validation performance (elastic net bordering on LASSO for red and ridge regression for blue). Both models found significant striking accuracy of the red fighter to be the most important variable for predicting the betting odds for red or blue, respectively. Finally, our LDA model to predict the method of victory for a fight performed especially well when predicting outcomes that ended in a decision but did poorly on fights that ended in a submission. This was likely due to there being so few fights that actually ended in a submission in our training data.

Reflection

One topic we wish we had been able to do was analyzing how different fighting styles and martial arts backgrounds matched up versus one another in UFC bouts. Unfortunately, this would have required an immense scraping effort on our parts since not much data on this topic was out there in a usable form. Additionally, anyone conducting this analysis would need to carefully consider how to break the data up to ensure we didn't have multi-valued attributes since fighting styles and martial arts training classifications are not mutually exclusive. We did not see this as a feasible topic given the time constraints of this class, but we hope others consider exploring this in future studies.

We mentioned earlier in the paper our concerns about the influence of outside factors on the gambling side of UFC and we wanted to elaborate on this further. In many cases, the betting odds of a fight are set based on how people have been betting so far which, in turn, was established by the starting betting lines. Gambling institutions adjust these betting lines so that, regardless of the fight's outcome, they break even in terms of their payouts and manage to get a profit from the commissions (additional fee on top of the amount being wagered which is required to place a bet) on the losing side's bets. Our analysis has shown that fighter data possesses some value for predicting the odds of a UFC matchup, but, counterintuitively, we learned in our research that the betting odds for a fight may not always reflect the actual odds of the fight ending in a certain result.

A future way to explore the classification problem for method of victory could include QDA models which we were unable to implement in our problem due to issues with the data. Support vector machines may also give valuable predictive models for this problem. Unfortunately, we did not have time to implement these techniques given the time constraints of the class and since we learned about support vector machines near the end of the semester.

Contributions

Taylor was responsible for answering the supervised learning question with betting odds. She used elastic net that used different values of alpha so she could see the efficacy of lasso and ridge and models that were in between. She was responsible for the write up regarding this question as well.

Nikhar used community detection to examine potential communities within the data with hopes of trying to find underlying structure. He also cleaned the initial datasets as there were a few errors in data such as cells being nulled out. Additionally, he wrote the section in the paper regarding community detection.

Jason used unsupervised clustering techniques (k-means and hierarchical) and then wrote that portion of the paper.

Michael helped clean the data set that contained the manner in which the fight was won (classification dependent variable). He was also responsible for the classification problem, and he did the write-up for that as well.

Drew cleaned the career averages dataset and merged the betting odds data with this dataset. Additionally, he helped each of the other members in their analysis when issues were encountered. He also wrote the introductory sections, the conclusion, and the reflection for the paper.