**Research and Background**

For this project I have decided to use UFC fighter data. I have been a big fan of MMA for over 8 years and have been watching UFC for just about as long. I thought it would be very interesting to go online and try to find a dataset that includes fighter statistics and try to make something meaningful out of it. My initial questions were to see if I can show what attributes are most important in a fighter, if I could predict who would win a fight between two fighters, and if I could predict who has champion potential. After getting my dataset I stuck with only the first two questions. So, the goal of this project is to show the most important attributes of a fighter, and to predict outcomes of fights.

**Dataset Description and Cleaning**

My first step was to import the data into Jupyter Notebooks as this is where I was going to do the cleaning. I then printed the null values and the shape of the dataset.

A screenshot of a computer

AI-generated content may be incorrect.

As you can see, my dataset has 18 columns and 4,111 rows. There was a good amount of null haves that needed to be dealt with.

The first column I went to fix was reach\_in\_cm. For this column I thought the best way to fill any null values would be to fill it with its corresponding height value. I knew this wouldn’t be exact, but I thought this would be the most accurate way to do it.



This is the code that I used to accomplish this task.

Next, I got rid of the nickname column all together as I saw no way of using it to my benefit.

A close-up of a logo

AI-generated content may be incorrect.

This is the code that I used to do that.

I wanted to create an age column because I believed that it would be a useful statistic for predicting fight winners and an important attribute. I later decided that the age column will not be used for my purpose because not all fighters on my list are current fighters. I kept it in just in case I wanted to filter my data to only current fighters.

A screenshot of a computer code

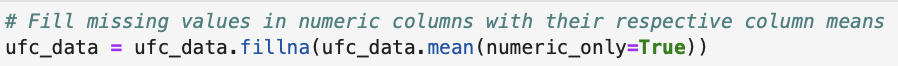
AI-generated content may be incorrect.

I used this code to accomplish this; after creating the age column I had to round it to the nearest integer to get a round number for my ages.

Next, I handled the stance column. For this column I decided it would be best to fill all null values with ‘Orthodox because 80% of all fighters use and orthodox stance.



After completing this I had 3 columns with remaining null values. My height, weight, and reach columns have null values left over. To deal with these I filled them with their respective column means.



Next, I wanted to create a couple of new columns to use for my models.

A screenshot of a computer code

AI-generated content may be incorrect.

I create a strike\_differential column, which is calculated by the number of strikes per minute someone is landing minus the amount they absorb per minute. I also created a win\_percentage column as I wanted to make use this in my model. This was accomplished by dividing wins by total fights. If these new fields created any null values, I just dropped them.

Next, I created a new column to show the weight class each fighter belongs to.

A screenshot of a computer program

AI-generated content may be incorrect.

For this I had to use continuous ranges to avoid having ‘unknown’ weigh classes in my data.

Then, I created new datasetsfor each weight class.

A screenshot of a computer program

AI-generated content may be incorrect.

Again, I used continuous ranges to avoid losing data. A tolerance was set to avoid minor discrepancies, but that is altogether avoided with the continuous ranges. After completing this I checked the shape of these new datasets to make sure that they had data, they did.

**SQL Queries and Outputs**

I used SQLight Online for this part of the project as I did not need to do advanced queries, and this was a very simple way of completing this task. My goal in SQL was to grab a few averages based on weight class to showcase differences in how the weight classes fight. First, I imported my dataset, and a table was automatically created for me. Then, I wanted to see the average significant strikes landed per minute for each weight class.

A computer screen with white text

AI-generated content may be incorrect.

This is my SQL code to complete this task.

A screenshot of a black screen

AI-generated content may be incorrect.

This was my output. This output is not in any specific order, but from this output we can see that the lower weight classes have a higher output on average. The highest output is Flyweight (the lowest) while the lowest is Heavyweight (the highest). This makes sense as it takes the heavier fighters more energy to throw punches.

Next, I wanted to look at the average striking accuracy for each weight class.

A screen shot of a computer code

AI-generated content may be incorrect.

This is the code used to find it.

A screenshot of a computer

AI-generated content may be incorrect.

Here is my output. We can see a very similar trend in accuracy that we do in number of strikes thrown. The lower weight classes have higher accuracy that the higher weight classes. Again, Flyweight has the highest, while Heavyweight has the lowest. This is not exactly the same as number of strikes as some weight classes do not follow this rule, like Light Heavyweight, but in general the trend is similar. I believe this is due to lighter fighters being able to stay sharp with stamina because they are not tiring out their arms as fast as the heavier fighters.

Finally, I wanted to see the average number of takedowns landed per 15 minutes by each weight class.

**A computer screen with white text

AI-generated content may be incorrect.**

Here is the code I used to do that.

A screenshot of a computer

AI-generated content may be incorrect.

This is my output. A similar trend is shown in this output as the others, but not nearly as different. A lot of the weight classes show very similar numbers. Again though, Heavyweight is showing the lowest output while Flyweight is showing the highest.

From these outputs I think it is safe to say that the lower the weight class is, the more activity there will be in the fights on average. This is most likely due to the amount of energy that is being consumed by the bigger fighters to do the same things the smaller fighters do.

**Python Code and Outputs**

For my model I tried many different options but the one I want with was a Gradient Boosting Model with K-fold Cross Validation

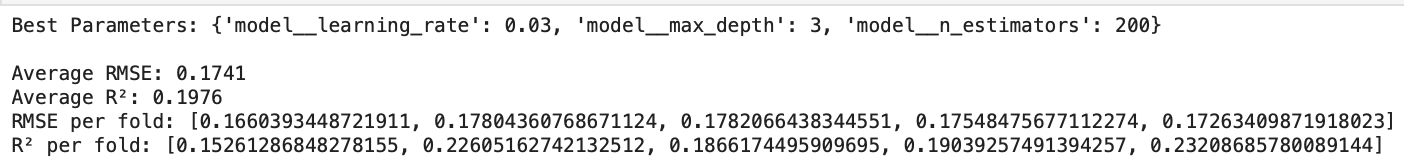
A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

I used this model to show what attributes are the most important when trying to figure out what causes a good win percentage for a fighter. There was a lot of tuning and tweaking done to this model as it was not predicting a good amount of variability in win percentage. I also created new features to use in this model to try and get the r-squared number up.



Overall, this model does not really work well when trying to show what attributes contribute the most to win percentage because it is only predicting about 20% of the variability in win percentage. This model was the best one that I could use to get that number up. There are so many other factors that contribute to winning a fight, and I don’t think it is something that can be predicted using statistics accurately. Sometimes, the better fighter does not always win because there is so much unpredictability in the sport.

For the sake of this project though, I will use these features and shed a little bit of light on what help win percentage.

A graph of blue squares

AI-generated content may be incorrect.

This is the output of my model. According to this model, the best way to improve your win rate is land a high number of significant strikes per minute. It is by far the most important feature in this model. The second most important thing is to have a low off strike ratio, which refers to the number of punches you are landing compared to the amount you are taking. It is not only important to throw many strikes, but you must also avoid being hit. Third, it is important to have a good height to reach ratio. You do not want your reach to be proportional to your height, the longer your reach, the easier time you will have fulfilling the other two conditions. Most of the more important factors have to do with striking, so this model tells us that over history UFC fighters have found more success with a good stand-up game than with a good ground game. That balance has been shifting in the sport as I have been watching it, and I am interested to see what this type of data would look like in 5 years.

Next, I build a model to predict the outcome of two fighters.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

This model was made using the flyweight dataset so all the fights being predicted are flyweight fights. I have made all the datasets for the other weight classes so that this can be interchanged if anyone wants to look at other weight classes, but for the purpose of this project, one weight class will be enough. This model has an accuracy of 74%.

A screenshot of a computer

AI-generated content may be incorrect.

These are fights that I simulated using this model. As you can see, this model did not work too well. First it predicted Brandon Moreno would beat Alexandre Pantoja, when in fact he has lost to him three times already. Next it predicted Brandon Moreno would beat Francisco Figueiredo with 72% confidence, when they have an even record against each other. The next fight never happened it was just for my own curiosity because Demetrious Johnson is the best fighter to ever be in that weight class so I wanted to see if it would say he wins. Next it said Brandon Moreno would beat Kai Kara-France, which is correct, but its only 54% confident which is lower than its wrong predictions. This model and this project have shown me that fights cannot be predicted based off statistics because there is just way too much uncertainty when it comes to fighting.

**Tableau Visualizations**

**A graph of blue bars

AI-generated content may be incorrect.**

**A graph of different sizes and colors

AI-generated content may be incorrect.**

These are 2 snippets of a dashboard I created with tableau. The goal of this dashboard is to showcase the difference in fighting styles by weight class. As we can see in the top right, The lower weight classes show much more activity with takedowns than the heavier ones, probably due to energy consumption. Same goes for submissions as well. This rule is not perfectly true though as Middleweight has the highest average submission attempts, showing that maybe that weight class has been grappler heavy for fighting styles. For strike differential it is all over the place with Featherweight being the worst and Light Heavyweight being the best. This shows that the Light Heavyweight weight class has a lot of talented strikers who can land punches at a higher rate than their opponent. Strikes absorbed follows that first rule, the lower classes have the most while the higher classes have the least. I believe this is due to the amount of power behind there strikes. It does not take many strikes from a Heavyweight to end a fight because there is so much power so it makes sense for them to have less.

Next, I made the same charts but for the different stances a fighter uses.

A comparison of a graph

AI-generated content may be incorrect.

A graph of different sizes and colors

AI-generated content may be incorrect.

First, we can see that the Southpaw stance attempts the most takedowns but is pretty like both Switch and Orthodox. However, both Open Stance and Sideways have very low or zero takedown attempts on average. For sideways stance I would assume this is because the amount of data in that column is low and maybe there is not any people with that stance who wrestle, for Open Stance I do not know why there are so few takedown attempts. Next, the Sideways stance shows the highest average submission attempts by a long shot. This must mean that the fighters in the stance have good ground game but prefer to not be the one to initiate the ground fight. Next, the Orthodox stance has the best strike differential, and there is a good-sized gap in between each stance. None of them are similar. Orthodox and Southpaw are both traditional boxing stances so it would make sense for them to be the best for striking. Finally, the Switch stance absorbs the greatest number of strikes on average which must mean that switching back and forth leaves you open to attacks and easier to land on in those times. Sideways stance must leave few openings for attacks, and most be a good stance for avoiding your opponent’s attacks.

**Conclusion**

From this project I feel like I have gained a better understanding of the components of a fighter, however I do not think this understanding is close to complete. I think with the data available for use it is not possible to truly dictate what makes a fighter good or bad and allows you to predict who will win fights. I think you would need so much data about intangible things, like a fighter’s mindset and mental strength to really get results. I am happy that I did this project with this data though because it is something that really interests me and now, I know a bit more about the data that I will need to do this again in a more accurate way.