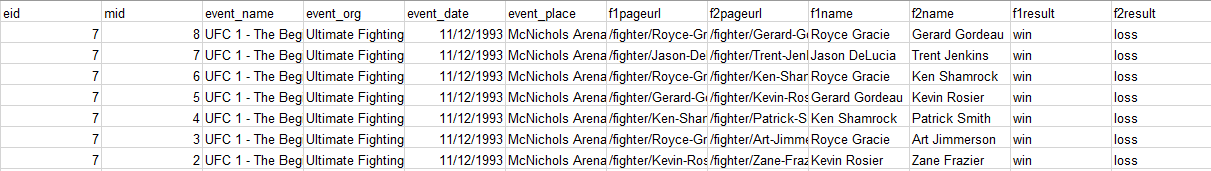
Taylor Moorman

IST 652

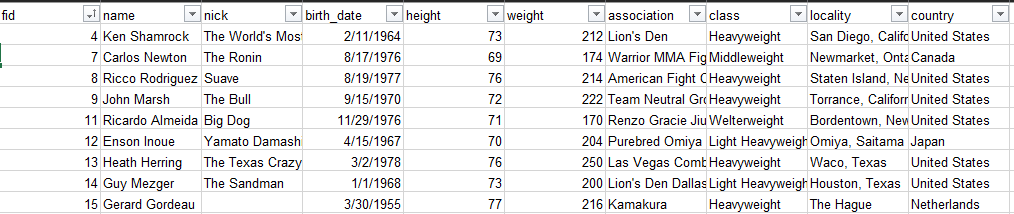
June 7, 2019

Final Project: Predicting UFC Fights

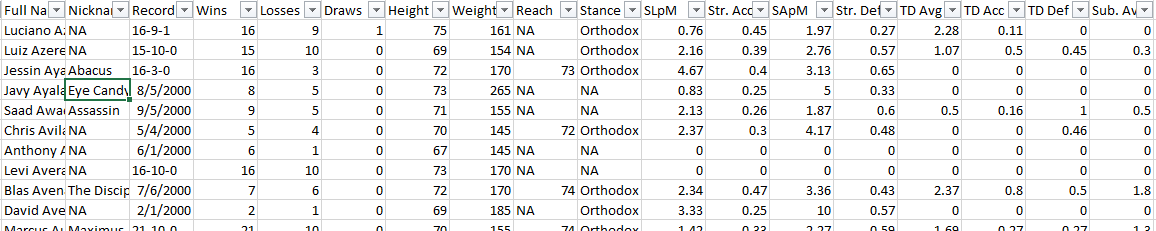
**Data and its source:**

 For the purposes of this analysis, three separate data sources were used. The first two datasets, called fights and fighters, are extracted from an mma site called Sherdog.com. This website is a main hub for all things mixed martial arts(mma) and Ultimate Fighting Championship (UFC). The fights dataset holds data from every UFC fight from the company’s inception in 1993 to 2016. Some of the fields in this set included the event name, location, fighters who fought, the referee, and how the fight ended, including the time it ended and by what means it ended (decision, knockout or submission). A snapshot of the data is included below:

The second dataset from Sherdog was a fighters set, of every fighter to fight in the UFC during the same time period as the fights dataset. This set contained information about the fighters like height, weight, nationality, gym association, birth date, and their names. A subset of that dataset is as follows:



The last dataset is from the site FightMetrics. This company does in depth analysis of fighters and fights and provides advanced metrics on fighters. This included not only their reach, but also their takedown accuracy, which is their successful takedown attempts over their total attempts. It also included strike accuracies, strikes landed per minute, defensive accuracies, etc. This dataset was important for being able to get an understanding of each fighters individual styles and getting a quantifiable variable of how “good” any fighter is at a specific area of the fight game.



All three datasets were collected in the form of excel documents.

**Data Transformation Steps (Preprocessing):**

The data transformation with these three data sets can be broken down into subsections:

1. Fights- First, the fights dataset was filtered to only contain fights that happened in 2003 or later. This was done because the UFC underwent large rule changes in the early 2000’s that made previous fights obsolete since the context surrounding the fights was so different. Next, the dataset was pared down to only a few of the columns. This was done because other columns were either superfluous, not needed for predictions, or contained junk data.
2. Fighters - a few junk columns of URLs to webpages were removed from the fighters dataset
3. Fightmetrics – This set was pared down to relevant columns, and then any fighters that had a null fid were removed. This was done in order to set up for the merging of data sets on fid later on.
4. Fighters was then merged with fightmetrics on fid in order to create one wider data set of any and all data pertaining to specific fighters. This action created dataframe fNEW.
5. Fights was broken into two separate sets of wins and losses. Previously, each row had been one fight, including both the winner and losing fighter, with no pure outcome column. The set was split apart into two dataframes of wins and losses and the appended to create a new dataframe of each row containing one result for one fighter in one given fight. This way, a testing variable could be made “y” where 1 is a win and 0 is a loss.
6. This totalfight set was then merged with the fNEW dataframe to create the finalset dataframe, containing all of the fight and fighter data. It is this set that will be modeled off of.
7. A number of fields were removed from the finalset that were not needed for predictions, and/or through trial and error, provided no help in improving model performance.
8. The field “stance” was converted into binaries where a 1 is orthodox and a 0 is southpaw, this was done to turn an object variable into an int. “Age” field was then created by subtracting the birthdate of a fighter from the event date to ascertain their age at the given fight.
9. Any and all remaining NAN’s were removed.
10. Lastly, the fields “country” and “ref” were encoded to create int type variables for predictions and the previous object type fields were removed from the dataset.

**Methods of Analysis:**

For this experiment, first, descriptive statistics using grouping and frequencies will be collected to show the various ways in which a fighter is more likely to win any given fight. These statistics will be based off of the features in the dataset. Next, a correlation matrix will be formulated to understand the effect that each feature has on the target variable, “y’. Finally, a RandomForest Classifier will be used to created a model to predict the outcome of the fights.

**Questions:**

The main question of this analysis is whether or not UFC fights can be predicted with any semblance of accuracy. “Semblance of Accuracy” will be defined by beating Vegas odds, or rather, can this model be better, than simply picking the favorite to win every time. This table from Oddshark.com, outlast the percent of the time that an underdog wins in a number of different sports.

| League | Underdog Win % |
| --- | --- |
| Major League Baseball | 42.3% |
| National Hockey League | 40.6% |
| Ultimate Fighting Championship | 34.5% |
| National Football League | 33.6% |
| National Basketball Association | 31.1% |
| NCAA Basketball (all divisions) | 25.7% |
| NCAA Football (all divisions) | 22.2% |

The most volatile sport seems to be Major League Baseball, with favorites winning only roughly 58% of the time. Therefore, it will be the goal to beat that standard, since inherently fighting is a much more volatile sport than the team sports listed in this table.

Secondly, aside from full-on prediction, what features are the most important in winning fights? Which ones correlate the most with winning?

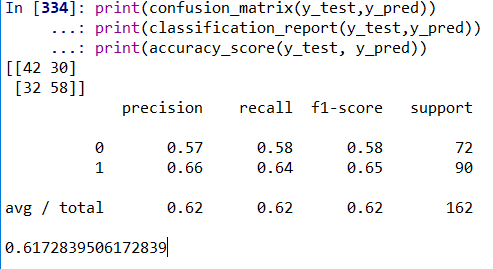
**Brief Description of the Program:**

The program written for this experiment takes in the aforementioned three data sources, transforms and cleans and merges them into one final data set. The program then creates and outputs descriptive visualizations and tables on the data, then trains and tests a RandomForest Classifier and outputs the results and associated performance metrics. Finally, the program creates and outputs a correlation matrix of all of the features.

**Conclusions:**

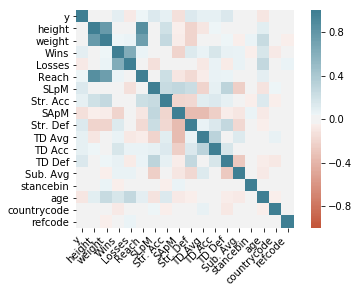
The results of this experiment were certainly a mixed bag. Although overall, both original goals/questions of this data were met.

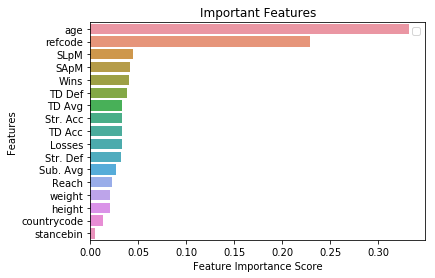
Can UFC fights be predicted in a fashion that beats Vegas odds? The answer to this question is yes, although admittedly this model is not good enough to be put to use to win money in actually betting on the fights. The original goal was to beat the betting favorites performance for Major League Baseball at 57.7% accuracy. The RandomForest model achieved an accuracy of 61.73%. with a confusion matrix and performance metrics as follows:



The model ending up being better at predicting losses than in predicting wins but did achieve the stated goal of beating Vegas Odds. This is important because as volatile as baseball is, the benchmark for this analysis, UFC fights are inherently much more volatile because of the violence and variety of ways to win.

The second question was: What features are the most important in winning a fight? The answer to this question is much more straightforward than in overall predictions. It is very clear from this analysis what attributes have the biggest impact on the outcome of any individual fight. From the correlation matrix and feature importance charts below, it can be inferred that a fighters age, the wins on their record, their Significant strikes Landed per Minute (SLpM), their Significant strikes Absorbed per Minute (SApM), and takedown defense (TD Def) are the most important factors that attribute to the outcome of a fight.





Intuitively, these variables make sense, Age can play a huge factor, whether a fighter is too young and hasn’t reached their athletic prime, or too old and no longer can physically perform the way that they used to. Next, the wins on their record follows because the more a fighter wins, the more likely it is that they are a good fighter, however at a certain point that can turn to a negative because more wins equals more total fights which can be taken as more overall damage to the body and an older fighter as well. Significant strikes landed per minute and absorbed per minute also make objective sense, because if you land more strikes you hurt your opponent more, and if you get hit more you are more likely to lose. Lastly, Takedown Defense is important because it means whether or not you can stop your opponent from dictating a fight, if you can stop them from doing that, odds are in your favor.

Overall this experiment yielded good results. The model, while beating the set standard, can’t totally be used for predictions because of its overall accuracy, the model and correlation matrix did provide a good look into the most important factors that go into determining the outcome of a fight. MMA and the UFC in general may be the most volatile and difficult to predict sport because it is impossible to quantify grit and determination and the heart of these athletes. An upset can happen at any time but looking into some of the most important factors outputted by this analysis can help give insight into future fight outcomes.