A statistical study of the Sweet Science

Sam Moradi-Balf   
City, University of London  
*sam.moradi-balf@city.ac.uk*

*Abstract*—This study aims to use data science techniques to analyse data about boxers to identify relationships between physical attributes and bout history, and to explore common views in the sport of boxing. The data for the project was obtained by web-scraping and consists of two datasets: one with boxer information and one with bout information. The study uses data visualization and statistical analysis techniques to explore the relationships in the data and potentially develop predictive models. The results of the study provide valuable insights for stakeholders in the boxing industry and may lead to the development of more complex predictive models. Specifically, the study aims to answer questions about the relationship between physical attributes and bout history with ranking. The results of the study suggest that there may be some relationships between these factors and can provide a starting point for further analysis and model development in the future.

Keywords—boxing, data science, statistics

# **Introduction**

“In any art, the prodigy presents a problem. Given too easy a problem, he gets slack, but asked too hard a question early, he becomes discouraged,” wrote A.J. Liebling in his book The Sweet Science, coining the phrase used to describe the pugilistic sport of boxing. The quote relates to the problem faced by boxing matchmakers when arranging bouts between respective boxers.

This project aims to use data science techniques to identify relationships existing between the physical attributes and bout history data of boxers. The goal is to identify whether relationships exist such as a boxer’s reach and their knockout ability, or whether certain divisions are more prone to knockouts than others. The business of boxing represents a multi-billion-dollar industry in the United States alone (Statista, 2014), with boxers repeatedly ranked among the highest earning athletes in the world (Forbes, 2022), creating a financial incentive to pursue accurate analysis and models for use by stakeholders within the industry.

The scope of this project is limited to data gathered from [www.boxrec.com](http://www.boxrec.com), specifically the top 50 male boxers in every weight category, however there are data such as the measurement of determination which are not as easy to score. Pic and Gudberg (2021) however found some success in applying T-Pattern detection techniques to identify boxing styles and the types of punches thrown which led to winning outcomes.

It should also be noted that advances in computer vision have led to a greater understanding of a boxer’s style through measurement of parameters such as their footwork, aggression, punch type, and output (Jabbr.ai). These advances could surely be incorporated into a predictive model to be used to predict the outcome of fights, while potentially shining light on some intangibles mentioned above such as the ability to take a punch.

# **Analytical questons**

This project seeks to answer is whether commonly held views in boxing such as “a good big boxer beats a good small boxer”, or that many boxers who climb the rankings have fought opponents with mostly “losing” records, have a basis in available data. We will also aim to answer questions such as the level of competitiveness in different divisions, as well as finding any correlations between different physical attributes and a boxer’s success as defined by their ranking.

These problems answered with the use of data science techniques may provide the opportunity to discover new insights into what makes a successful boxer, but also could provide the grounds for building greater and more complex models with which to assess and predict the success of boxers in competition.

The questions we seek to answer are as follows and will be looked at on a division-by-division basis:

1. For boxers in the same weight category, does a greater height and reach offer an advantage, as measured by divisional ranking and the number of wins?
2. Is it common for boxers within the data to have “padded” records, that is they have fought opponents with significant numbers of losses?
3. In which division is a knockout the more likely outcome of a bout?

Through addressing these questions, we will have explored the available data and created workflows for wider and more comprehensive, as well as providing valuable insights useful to stakeholders within the boxing community.

# **Data (materials)**

Well-maintained and compiled datasets for boxing proved difficult to find and so for this project I first began by scraping the unofficially but widely accepted best resource for boxing records, [www.boxrec.com](http://www.boxrec.com). I used web-scraping methods to obtain data on the top 50 male boxers of each division from basis information such as name, height, reach, and age, to information such as their boxing records, their opponents, and their opponents’ boxing histories.

This led to creating two separate data sets, one with boxer information, and one with the respective boxers’ bout information. It should be noted that there proved to be missing data within the records. In the first dataset there was not available data for all boxers with some missing values such as their height or reach. These issues were addressed using statistical methods and will be discussed further in this paper.

The first data set contained 793 rows and 18 columns of data, while the second data set contained 17,658 rows and 9 columns of data. There are a mix of data types which are wrangled with to make useable, as well as the use of encoding for some characteristics.

The data within the datasets permits the use of statistical methods to answer the proposed questions of this paper, as well as the ability to engineer further features or impute values from the data. I also intend to train a classification model using a Random Forest Classifier to see whether I can accurately predict the ranking of a boxer, based off the datasets.

While I believe the key characteristics of the datasets are useful, a limiting factor is the overall sample size. Due to some missing values, data are likely to be dropped. A broader sample size perhaps covering more boxers would yield better results.

# **Analysis**

## Data preparation

Without any easily accessible datasets, I had to develop a script to scrape boxing data from the web. This required the implementation of several python libraries to scrape the data, but also heavily required the use of data-cleaning methods to prepare the unclean data for future work. I first created a script to obtain the URLs of the desired boxers, those that met the criteria set out earlier before new scripts were written to create the two datasets that would ultimately be used.

Due to the structure and design of the webpages being scraped, there were many inconsistencies between pages making the programmatic scraping more difficult and requiring several refinements of the scripts to become successful. One function I created which became useful in both scripts was that to clean the name before saving to another file. This used the regular expressions (RE) module as well as the built-in replace() function to clear unwanted HTML code and return a clean name. Once all the required pages had been scraped by each script, the data was batched together and formalised in respective Pandas dataframe structures and saved as two comma-separated-value files.

## Data derivation

Once the datasets were ready, I imported them into a Jupyter Notebook and began to look for ways to generate interesting new features. I first began to sort the fighters by division, and their ranking within their respective division. I encoded the different divisions using integers between 0 and 15, with 0 being heavyweights and 15 being minimumweights.

Here I tackled an initial issue with my datasets, in that some key metrics such as height, reach, and age, where missing from some of the boxers’ data. Here, I decided that I would not be able to reliably impute age considering the characteristics existing within my dataset, so I dropped those missing values, while I used the mean height value for boxers in respective divisions to fill any missing height values in those divisions. Once complete, I then calculated the difference between height and reach for every boxer where both values existed, before again calculating an average difference by division. Finally, I added the existing height value to the new average height and reach difference by division value to impute a reach value for those missing values.

A smaller number of rows were removed due to missing birthplace or stance data. I made attempts to train a Random Forest model to predict reach values given height values, however the small size of the sample (overall 683 rows before training and testing split) left me with a very low accuracy score, and so I decided to stick with the mean-derived values.

I also engineered another feature to score a boxer’s quality of opponents. The formula for this score is as below, where oW denotes the sum of the boxer’s opponents’ wins, oD denotes the sum of the opponents’ draws, and oL denotes the sum of the opponents’ losses.

oQuality = (2\*(oW + oD)) / 2 \*(oW + oL + oD)

This provides a numerical score between 0 and 1, with a score nearer 0 implying the boxer’s opponents have significantly losing records overall, while a score nearer 1 implies the boxer’s opponents have generally few losses and so represent stronger opposition.

## Construction of models

Following the engineering of features, I decided to use K means clustering to see whether any patterns would emerge from the data. First, I created a new dataframe with a group of characteristics that I felt most appropriate for the model, then I began by deriving silhouette scores for K values between 2 and 10 for the dataframe and noted that a K value of 6 would yield the best result identified by the mean silhouette score across K values.

Chart, line chart

Description automatically generated

I then proceeded to construct a K means clustering model using a K value of 6, which was applied to the dataframe and plotted the results of the model, as well as saving the cluster label for each boxer in a new column of the dataframe for future reference. From these plots I noticed an interesting pattern where there seemed to be more outliers in terms of height and reach in “super” divisions. These are those such as super welterweight, which lies between middleweight and welterweight. I decided to run a search for the number of outliers, using the interquartile range method of identifying outliers, however the results showed very few outliers within the data.

A Random Forest model was also constructed to attempt to predict the division rating of a boxer based off the characteristics present within the data, however this yielded a low accuracy score. Attempts to tune hyperparameters with varying values of trees in the random forest and the maximum depth of the trees still yielded poor results.

## Validation of results

The poor accuracy of the Random Forest models I attempted to use are likely down to the sample size of the data, and perhaps it may have been better to expand the scope of the data as previously stated.

Further, the K means model showed varying optimal K scores for respective divisions, however I chose to use a single K score for all the divisions within the dataset, which may have reduced the accuracy of the model. Here, it may have been better to train models for specific weight categories, and again use a much-expanded dataset of boxers per division.

Some results were similar to those observed by Sorowski (2014) where stance showed no real bearing on performance.

# **findings and reflections**

The initial analysis undertaken showed some logical inferences, firstly that on a young-to-old scale, boxers within the dataset are right-skewed, with most boxers below 31 years old. Further analysis showed that boxers in the lower weight divisions tend to be younger, while those in the heaviest divisions are of the oldest age.

Chart, histogram

Description automatically generatedAnalysis showed that boxers generally tend to have a greater reach than height, while height and reach decrease as the weight of boxers decreases.

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

Looking at the average of round fought per division, a weak trend was observed which showed a tendency for boxers in the lower weight divisions to box more rounds, than boxers in the heavier divisions. This trend was reversed for the average knockouts per division, with boxers in the heavyweight division exhibiting a greater knockout percentage than those in the minimumweight category. Here, we can see an answer to one of the questions posed in that the data are highlighting a relationship where heavyweight boxers tend to fight less rounds with a significant higher knockout percentage than those in weight categories below them.

From this we can also move to including some engineered featured such as the average opposition quality score, as well as the average of the net balance of wins and losses for boxers by division, to identify whether there is a greater prevalence of ‘record padding’ in any weight classes compared to others. Here, the data are indicating that the mean opposition quality score is generally between 0.6 and 0.7. However, between these scores there is an interesting inference that the cruiserweight division (div index = 1) has the lowest score, but the lower weight divisions generally tend to have both low net wins-to-losses as well as quality scores on the lower end.

Chart, line chart

Description automatically generatedThis prompted a further analysis instead taking the median score of the division, which this time pointed to light heavyweights having the lowest quality score, but also a clear downwards trend in net wins and quality scores as the weight of boxers decreases. Further, by incorporating opponent losses, we can see that cruiserweight boxers have the lowest quality scores and highest opponent loss, indicating perhaps their records are more padded.Chart, line chart

Description automatically generated

By creating linear regression models on the data for height and against divisional ratings by division, we cannot reach a clear conclusion to answer our final question.

Here, we can see that in different divisions, the height of a boxer is not a clear indicator of their division rating. P-values were broadly weak except for in the light flyweight division where height showed greater significance. Some divisions such as the super middleweight division showed steeper slopes than other divisions, and this is likely due to skew because of boxers with below average height for the division having low division rating values (a higher ranking).

In other divisions such as the middleweight division or super welterweight division, slopes were very flat indicating that most boxers in fall within a relatively clear boundary of height values, however from plots show some clear outliers within the data.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

In hindsight, it may have been possible to use the data collected to search for bouts between boxers and encode a win, loss, or draw value with which we could include into the models, and which could add greater context.

Further, having the dates of the bouts would permit to us to see how long boxers fought opponents with low quality scores.

Overall, the project would have benefitted from a greater volume of data which would have improved the accuracy of the classification models and led to more reliable results.

**Word counts**

Abstract = 149 words

Introduction = 295 words

Analytical questions = 243 words

Data materials = 296 words

Analysis = 904 words

Findings and reflections = 596 words

Total word count = 2,483 words.

**References**

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. (Statista, 2022) - Statista Research Department, November 4 2022, https://www.statista.com/topics/10128/boxing/
2. (Forbes, 2022) - Knight, Brett, (May, 5, 2022)https://www.forbes.com/sites/brettknight/2022/05/11/the-worlds-10-highest-paid-athletes-2022/?sh=7342bacf1f6c
3. (Jabbr.ai) - www.jabbr.ai
4. (P Sorokowski, 2014) - Sorokowski, P., Sabiniewicz, A. and Wacewicz, S. (2014) “The influence of the boxing stance on performance in professional boxers”, Anthropological Review, 77(3), pp. 347–353. doi: 10.2478/anre-2014-0025.
5. Pic and Gudberg (2021) - Pic, M., Gudberg K., J, (2021) "Professional boxing analysis with T-Patterns), Physiology & Behavior, Volume 232, doi:10.1016/j.physbeh.2021.113329.