Shane Viola Data Mining CIS 9660

The goal of this model was to develop a machine learning model to predict the outcomes of UFC fights based on fighter statistics. Predicting the outcome of a fight is often very difficult and even sports betting sites have an average accuracy between 56-58%. My data was sourced from Kaggle and has every fight from 2010-2024. I first checked all the data and decided what columns to keep and what columns to get rid of. All columns related to post fight aftermath; betting odds and low correlating features were removed. I then created new features that calculated the difference between two fighters’ attributes, like take downs landed difference or significant strike accuracy difference, or difference in reach. For each two fighters they were labeled red fighter and blue Fighter. I noticed there was an imbalance between wins for red Fighter that can affect the outcome of the model, so I flipped the data so that it was balanced wins for both colors.

After cleaning the data, I made a chart that plotted a features correlation with the Red Fighter winning. I was surprised by the outcome, the biggest correlated features were the age of the opponent, the win/loss ratio of the red fighter and the average takedowns landed for the red fighter. I made a correlation matrix of the features and made a bar chart for method victory. After splitting my training and testing data 70/30 I did feature scaling and ran 7 machine learning algorithms. I found Support Vector Machines to be the best algorithm with a cross-validation score round to 60% on test data. There was also a small difference between training and testing data which meant it was not being overfit. In many iterations of feature selection, random forest was more accurate, but there was a high difference between testing and training data. K-Means performed the worst and was only slightly more accurate than a coin toss with an overall accuracy of 51%.

Completing this project, I learned a lot about how to optimize machine learning algorithms. I had to do a lot of trial and error with choosing what features to include as many features that I figured would matter did not and many features that I don’t consider had a surprising correlation with winning. Before I made this model, I used to bet on UFC fights and mainly base my bets on the number of wins and losses each fighter has as well as the number of knockouts. However, after completing this project, I found that the number of knockouts a fighter has or even the number of wins does not have a significant correlation with winning a fight and it was so low that I did not use past methods of victory as a feature in my model. Most importantly I learned from this project to have an open mind as you can have initial assumptions that turn out to not be backed by the data, and the value a machine learning model can provide.

In terms of real-world applications for this model, it provides fight analyst quantifiable data to access a matchup. In betting, it can enhance the decision making by giving a quantifiable likelihood of each fighter's victory. Anecdotally, I have shared the website with a few of my friends, and some have used it to bet on UFC fights that occurred on August 9th, 2025.

Appendix

A graph with blue and white lines

AI-generated content may be incorrect.

A screen shot of a chart

AI-generated content may be incorrect.

A graph of a graph showing different colored squares

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.

A graph with a line going up

AI-generated content may be incorrect.

A graph of a graph showing a number of different types of objects

AI-generated content may be incorrect.

A blue squares with white text

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

Citations

For visualizations ChatGPT was used to create charts in MATPLOTLIB.

In addition, some charts were sourced from examples presented in class or listed as requirements for the project.

[Ultimate UFC Dataset](https://www.kaggle.com/datasets/mdabbert/ultimate-ufc-dataset) sourced from Kaggle.