# UFC Fight Prediction Model

Sophia Manodori Greta Anesko

## Introduction

The Ultimate Fighting Championship is a multi billion dollar corporation that hosts fighting events around the world. International mixed martial arts athletes meet at these weekly events to fight for the division titles. Each match between two fighters has either 3 or 5 5-minute rounds and ends via submission, knockout, technical knockout (called by the referee), or by judge's decision. Our objective is to produce a classification model that can predict the winner of upcoming UFC fights, given the differences in fighter statistics between each fighter, with a low error rate.

## Data

The data is from the UFC Complete Dataset on kaggle that is parsed from the UFCStats website. It has detailed information on all 7226 UFC fights from 1996-2024, and contains all individual fighter statistics as well as the in fight statistics. The dataset contains 95 variables. For the purposes of this project, we will be primarily focusing on the fighter statistics as predictor variables for fight outcomes. Fighter statistics include physical statistics such as height, weight, and stance, and career statistics such as strikes accuracy and takedown accuracy. This dataset also includes many differences in fighter statistics between the Red and Blue fighters, all of which are calculated as Red minus Blue. We selected for these variables of differences in statistics for our model as the difference between statistics between the fighters is more meaningful for the outcome of the fight than any one fighter's isolated statistic, as the fight operates on the combination of skills of both fighters.

## Model

We performed variable selection on the differences in statistics variables using a logistic regression model to identify the most significant predictors. The variables we identified as significant predictors were the differences between fighters in: number of wins, number of losses, weight,

age, strikes landed per minute, strikes absorbed per minute, significant striking rate, takedown defense rate, and average number of takedowns per 15 minutes. We then split the data into a training set (the first 80% of the data) and the testing set (the most recent 20%). We tested four classification models: logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and Naive Bayes. As no predictors were highly correlated with each other except for differences in wins and differences in losses between fighters, there was reasonable evidence that logistic regression and Naive Bayes could perform well. As most predictors seemed to be relatively normally distributed within each class, there was reasonable evidence that LDA and QDA could perform well. We trained logistic regression, LDA, QDA and Naive Bayes models on the training set, then tested them on the test set and compared the error rates of each model between classes and overall error rates.

## Results

All models performed well, with overall error rates below 30% for each model. By comparing error rates between all four models, we conclude that although logistic regression and LDA have the same overall error rate of 26% and error rate for the Red class of 24%, logistic regression performs better by 1 percentage point for the Blue class, with an error rate of 38%. QDA had an overall error rate of 28% and Naive Bayes had an overall error rate of 29%. All models do substantially better at predicting the Red winner than at predicting the Blue winner. This is likely due to the majority of the wins being Red winners, as the Red side is usually the champion's side, so the greater number of Red wins allowed the models to better understand and predict Red wins. To test our model on upcoming fighting events, we compiled the fighter statistics for the UFC Fight Night on April 28 and ran the model on the fights. The model correctly predicted the outcome of 9 out of 10 fights.

## Discussion

A shortcoming of our model is its comparative weakness at predicting Blue wins compared to Red wins and its overall error rate being more reflective of its ability to predict Red wins. As we were designing a model that could best predict results of upcoming UFC fights, it made the most sense to create the test set from the most recent 20% of data; yet the true proportion of the test set was 89% Red wins, compared to the full dataset's true proportion of around 60% Red wins. This affected our model's ability to predict Blue wins. A future model that was less concerned with upcoming UFC fight accuracy might consider creating a training set and test set with proportions of Red and Blue wins reflective of the true proportion.

## Conclusion

Our model's significantly low error rates for each class makes it stand out from previous UFC fight prediction models. Our model is optimal for sports betting as it offers specific probabilities for each fighter, making it optimal for placing parlay bets since the risk can be minimized by choosing the highest probability fighters. It also occasionally will differ from the UFC favorite, highlighting ideal candidates for underdog bets.