MGT 6203 Project Final Report – Predicting UFC Fights with Machine Learning Models

Group # 45: Carson Stone (cstone305), Daniel Hardiman (dhardiman6), Divya Dronamraju (ddronamraju3), Stephanie Cortes (scortes30)

Contents

Introduction

This final report describes the entire process used to build models for predicting Ultimate Fighting Championship (UFC) fights results. This predictive modeling is intended to be used as a tool to help with betting on fight outcomes. To access the data and relevant code, please go to our team's Github website here: https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/Team-45

Background

The UFC (Ultimate Fighting Championship) is the largest promoter of mixed martial arts (MMA) in the United States and features some of the most talented fighters in the world. As the sport has risen in popularity, so has sports betting. Sports betting is a $155 billion industry and fighting ranks among the most popular sports in the betting industry (McQuaide, 2019). Each fight has odds assigned to it based on how the sportsbooks predict the fight will turn out. Beating the books and turning a profit betting on MMA has proven to be challenging, and therefore any insights/tools that could help gamblers gain an edge have the potential to be very valuable.

Similar work has been done in the past to predict MMA outcomes by using fighter statistics such as historical win/loss record and physiological measurements including height and reach (McQuaide, 2019). In this project, we will perform testing to assess if a similar model can be applied to UFC, and we will consider additional variables as well such as fighter fight movements during historical games, weight, stance, methods, etc.

Objective

The team's primary objective will be to build a model for predicting the outcomes of future fights using historical data from past UFC event results, fights, and players' stats. To achieve this objective, the team will first answer the Primary Research Question "Can we build a model that will predict the outcome of UFC fights with more accuracy than the traditional betting odds?" and the three supporting questions:

- Which features impact fight results the most?
- How accurately can we predict the outcome?
- Are there any outliers in the data? How could/should this information be used to help coach fighters or teams?

Initial Hypothesis

We anticipate that a reasonable accurate model (> 75% accuracy) can be built to predict the outcome of UFC fights. We expect that historical fights' results and physiological measurements will be the most significant variables to predict the outcome. If the fighter has won multiple times in the past, he/she has more experience to perform better, and if the fighter is heavier and/or taller, it might be easier for him/her to dominate the other fighter.

Methodology

The project will be developed in 3 main phases: data preparation & cleaning, modeling, and results comparison. The objective of the data preparation phase is to create an input table to use for training and testing various classification models. The dependent variable will be fight outcome and the independent variables will include fighters' historical fight performance, fights' statistics, and fighters' physical attributes and background. More details on the data preparation tasks will be described in the "Data Processing" section.

Once the input table is ready, the team will run a feature analysis model to select the variables that are significant to predicting fight outcome. This will reduce the number of variables used on the second phase of modeling classification models. Once the classification models are tested, the team will compare the models' results using confusion matrixes on phase 3 for easier interpretation and analysis. We will also be running separate models including the fight stats that are available after a fight.

To see the dependencies of the different tasks of the project see the "Approach Flow Diagram":

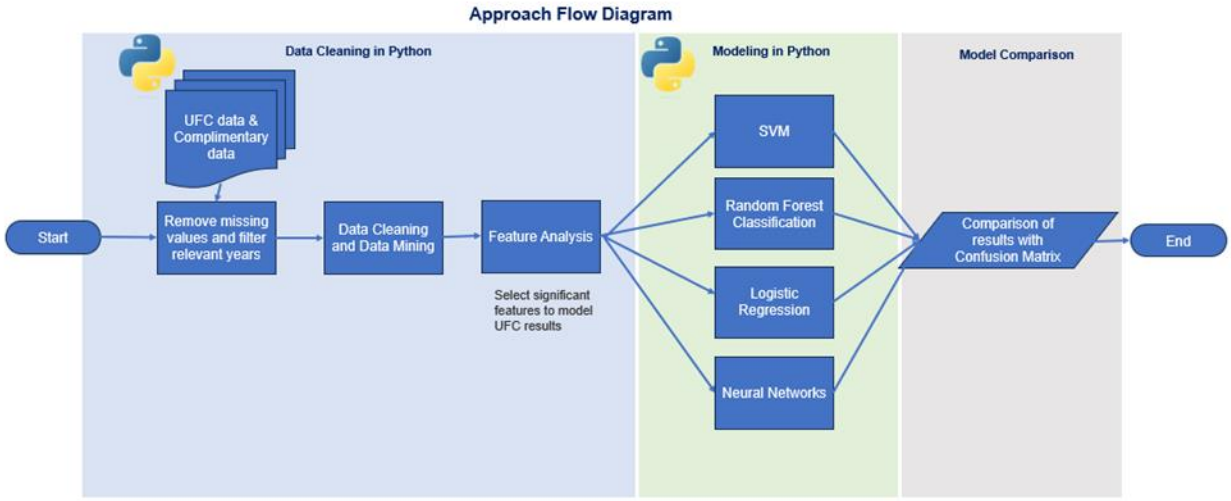


Figure 1: Approach Flow Diagram

Data Sources (GitHub link)

This repository by Grecco1899, found in our early research, provided all the project's raw data. This project houses six datasets (described below) that are updated weekly. The data is scraped directly from ufcstats.com and organized neatly into csv files. Ufcstats.com provides historical and real-time UFC fight data, and is officially licensed by the UFC. Some of the fight data in the csv files were cross-referenced to ufcstats.com and other sports reference sites like ESPN.com and sports-statistics.ufc.com.

Data Cleaning

This section describes the data cleaning process of the raw UFC datasets. The purpose of data cleaning was to transform the columns of the raw UFC data into a structured and data processing-ready format. Example steps in the data cleaning process included column standardization, data type conversions, and conditional column creation. Each of the raw datasets was cleaned and saved separately. For example, ufc_fighter_tott.csv has a cleaned version ufc_fighter_tott_clean.csv, ufc_event_details.csv has ufc_event_details_clean.csv, etc. Below is a brief description of the data cleaning for each raw dataset.

UFC Fighter Total Stats (ufc_fighter_tott.csv): Github link

We set out to clean and standardize the UFC fighter total stats dataset to make it ready for deep analysis. The process included simplifying column names, removing unnecessary 'weight' data, and converting 'height' and 'reach' to numerical values. We also transformed birth dates into a more usable format and extracted fighter IDs from URLs, making it easier to track and compare fighters.

UFC Event Details (ufc_event_details.csv): Github link

With the UFC event details dataset, our goal was to make the data more accessible and useful for examining event-specific aspects. This involved organizing column names consistently, extracting event IDs, and turning date strings into a uniform format. We also broke down the location information into city, state, and country to provide clearer insights into event locations.

UFC Fighter Details (ufc_fighter_details.csv): Github link

The focus for the UFC fighter details dataset was on ensuring the data was clean and straightforward to use. We made column names uniform and extracted fighter IDs from the given URLs, which helps in easily identifying and referencing fighters across different data sets.

UFC Fight Stats (ufc_fight_stats.csv): Github link

Our aim with the UFC fight stats dataset was to arrange it in a way that was conducive to detailed analysis. This meant making column names uniform, transforming 'ctrl' time data into total seconds, and breaking down stats columns to show 'thrown' and 'landed' figures separately. We also removed percentage columns to keep the focus on the raw numbers, which are more informative for performance assessment.

UFC Fight Results (ufc_fight_results.csv): Github link

For the UFC fight results dataset, we wanted to ensure the data was clean and structured for in-depth result analysis. We made the column names uniform, pulled out unique fight IDs, and standardized the outcomes, methods, and weight classes. Splitting the 'bout' column into 'fighter1' and 'fighter2' allows for straightforward comparison between fighters, making the data more intuitive to use.

Data Processing

This section describes the intermediate data processing steps of clean UFC datasets. The purpose of this section is to create new datasets by filtering, aggregating, combining, and further processing the clean datasets.

UFC Fight Results with Date (ufc_fight_results_date.csv): Github link

The dataset creation involved a methodical approach where UFC fight results were filtered by weight classes and outcomes to ensure completeness and relevance. Subsequently, event details post-January 1, 2010, were selected to focus on more recent UFC events. By merging these two datasets, a unified and detailed dataset was formed, capturing essential aspects of modern UFC fights, enabling focused and meaningful analyses.

Fighter Records Table: (fighter_records.csv) Github link

This is a transition table created to be joined to the cleaned features table (below). This was compiled from the Fight_results_with_Date.csv (above) by pivoting on fighter and summing their fight results. This gives us an important variable to establish how “good” each fighter has been, by recording their wins, losses, and total fights.

Features Table NOT including Fight Stats (cleaned_features.csv): Github link

The “UFC Fight results with date” table was cleaned by creating an “outcome” column with values of 0 and 1 to be used as the dependent variable (“1” --> fighter 1 won, “0” --> fighter 2 won). The dependent

variables were merged so that variables with suffix "f1" were going to include the player characteristics for fighter 1, and suffix "f2" variables were going to include the equivalent dependent variables for fighter 2. This table will be used as an initial test to select the significant variables; however, depending on the results from Lasso Regression or Logarithmic regression mentioned, the table might end up needing some additional cleaning before running the models.

Feature Table Not including Fight Stats with Lasso selection (cleaned_features_Lasso.csv): Github link

This table builds off the features tables (above). Then this was run through LASSO regression to obtain a smaller set of variables. Anything with a lasso value of >= .001 was kept as being potentially significant. This table leaves flexibility for more variable selection during the modeling phase if needed.

Features Table including Fight Stats (features_with_fight_stats.csv) Github link

This table builds off the features tables by adding in fight statistics to the list of factors.

Feature Table including Fight Stats with Lasso selection (features_with_fight_stats_lasso.csv) Github link

This table builds off the features tables by adding in fight statistics to the list of factors. Then this was run through LASSO regression to obtain a smaller set of variables.

Models to predict outcome of fights

All Variables included (Github link)

Model Name	Accuracy	Sensitivity	Specificity
Logistic Regression	82.3	85.9	77.6
SVM	81.9	85.6	77.1
Random Forest	81.5	86.9	74.5
Neural Networks	82.4	84.8	79.2
Voting Classifier	82.4	85.6	78.4

Lasso Selected Variables (Github link)

Model Name	Accuracy	Sensitivity	Specificity
Logistic Regression	72.4	76.9	66.4
SVM	72.6	77.4	66.3
Random Forest	73.2	75.5	70.2
Neural Networks	74.4	79.9	67.1
Voting Classifier	72.9	74.1	71.5

Model Interpretation

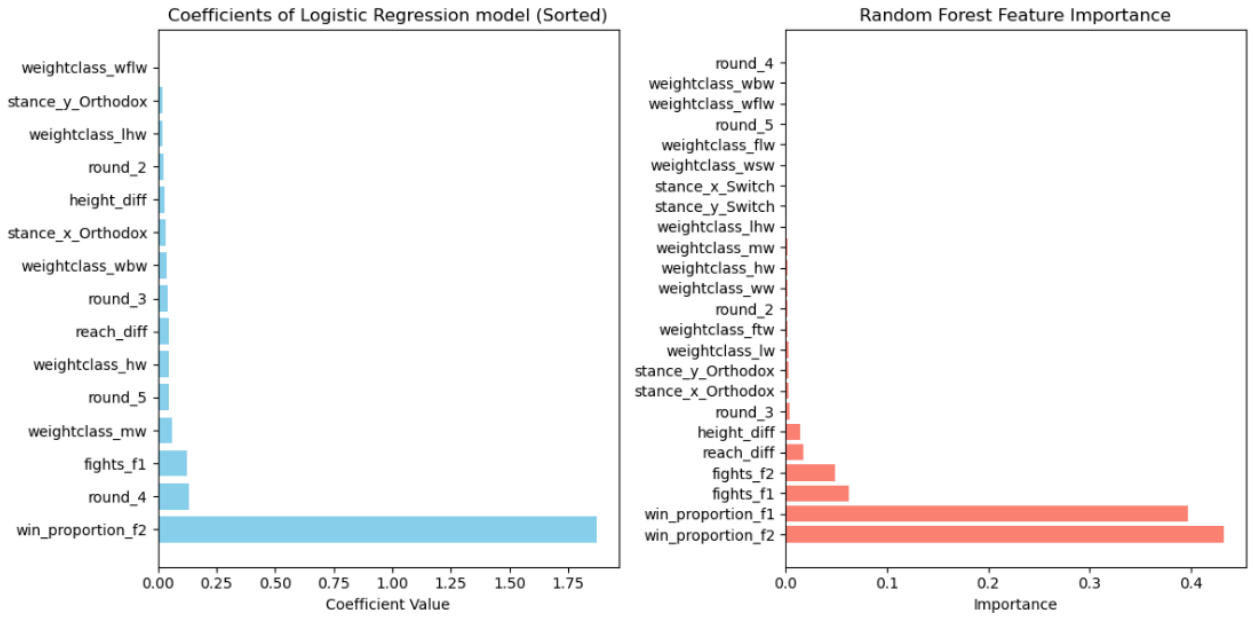
We created two sets of models, one including all variables i.e. 24 variables and the other including only 4 variables selected via Lasso regression. We did not include any variables that will not be available at the

time of prediction like time or any fight statistics like control (time a player remained on top of the opponent after a throw), significant throws etc. This data is available in the fight statistics dataset.

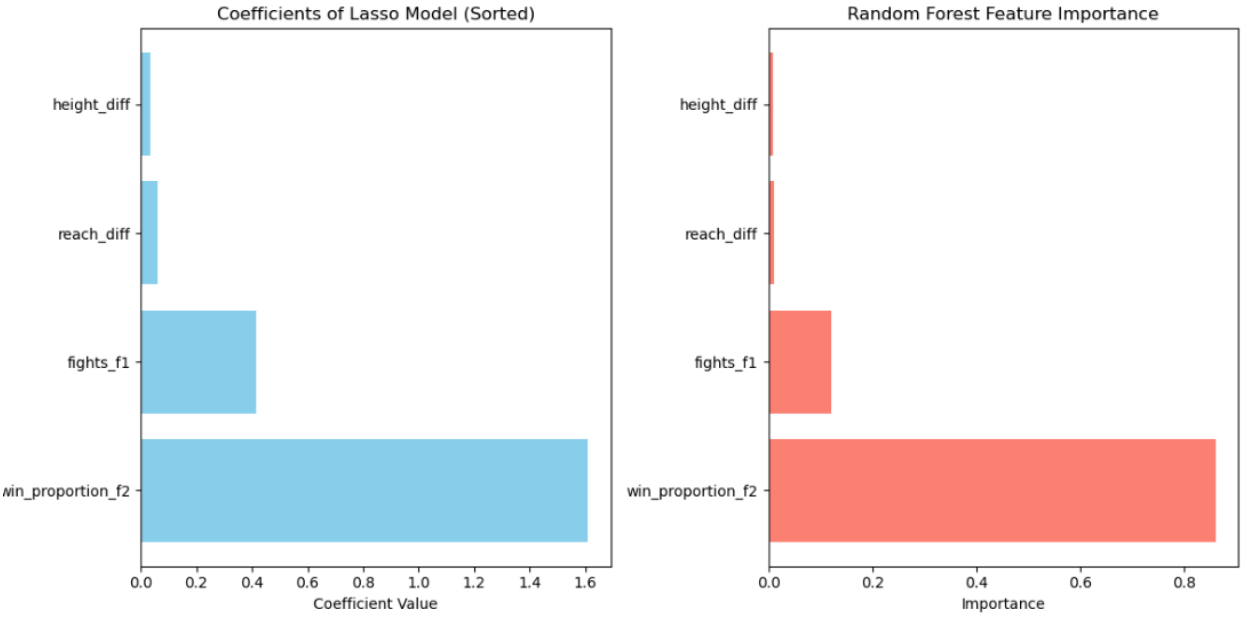
We observed that the Lasso model was very aggressive in slashing out the variables and ultimately reduced to a mere 4 variables: Total number of fights fighter 1 was involved in, Proportion of wins of fighter 2, height difference and reach difference between the fighters. This oversimplified the models and introduced bias in the model which caused underfitting. The voting classifier which took a majority vote of all the models produced an accuracy of 72.9%

Feature importance using Logistic Regression and Random Forest:

All Variables



Lasso Selected Variables



Our recommendation would be to keep all the variables to pick up the finer information available in all the variables. Among the four models generated: Logistic Regression, SVM, Random Forest and Neural Networks, the test data accuracy, sensitivity and specificity did not vary very much. Hence, we recommend using a Voting classifier that takes a majority vote among all the models and produces better and more reliable results.

Further details on recommended Voting Classifier

Confusion Matrix

	0	1
0	511	141
1	123	719

Precision Score: 83.6

Recall Score: 85.4

Sensitivity: 85.4

Specificity: 78.4

Test Accuracy: 82.3

Based on the metrics, we can say that the model is not overfitting to any specific class as the precision and recall are in a similar range.

In the model building process 5- fold cross validation and hyperparameter tuning was employed to ensure there was no overfitting to the data. As of the latest run, below are the details of the parameters selected that were ultimately used in each of the final models.

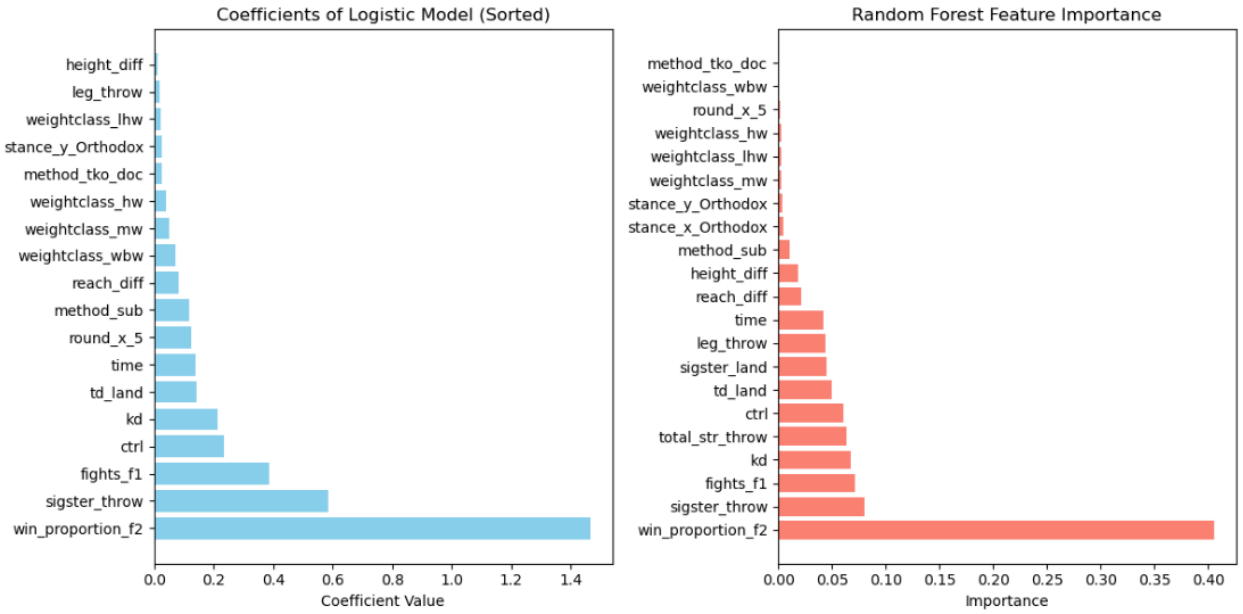
```
Searching for best Logistic Regression...
Fitting 5 folds for each of 7 candidates, totalling 35 fits
Best parameters for Logistic Regression: {'penalty': 'l2', 'C': 0.01}
Best accuracy for Logistic Regression: 0.8283039792872573

Searching for best SVM...
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best parameters for SVM: {'kernel': 'linear', 'C': 1.0}
Best accuracy for SVM: 0.8268704958854862

Searching for best Random Forest Classifier...
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best parameters for Random Forest Classifier: {'bootstrap': True, 'max_depth': 24, 'max_features': 'log2', 'min_samples_leaf': 19, 'min_samples_split': 1
2, 'n_estimators': 558}
Best accuracy for Random Forest Classifier: 0.8225655106449645

Searching for best MLP Classifier...
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best parameters for MLP Classifier: {'learning_rate': 'invscaling', 'hidden_layer_sizes': (100, 50, 100), 'alpha': 0.001, 'activation': 'logistic'}
Best accuracy for MLP Classifier: 0.8262966077936641
```

Traditionally, sports betting odds are best expressed as probabilities (e.g. Team A has a 71% chance of beating Team B). However, binary classification proved useful in this case since most fights have a clear favorite and there are rarely huge upsets. One thing that this model (as well as other models tested here) won't account for is ties. Ties wouldn't necessarily work with a model like this, unless there was some predefined range of values around the 50% mark that defaulted to a tie as a third possible outcome.

However, ties typically don't play a major role in combat sports betting, so excluding them from the scope of these models seems reasonable.

Models for Post-Game Analytics (Github link)

The second set of models was built solely for interpretation of fight stats, so accuracy is of less importance. In this situation, the models built after aggressively selecting variables via Lasso regression produced a less noisy and clearer picture. Below are the feature importances based on fight as well as player statistics.



Future Research

One area that seems promising for future research would be to experiment with using different sets of calculated predictors to train and test the models. Among the predictors in the cleaned dataset that we used for modeling, by far the most influential was each fighter's previous win/loss record (i.e. win percentage). While the models trained on this dataset ended up performing well, gamblers likely wouldn't find them terribly insightful if their recommendation was always to bet on the fighter with the better prior record. Furthermore, many of the factors in the raw data available to us such as the stats from the fights themselves (e.g. number of strikes, number of takedowns, number of submission attempts) aren't directly useful for predicting the outcome of a given fight since we wouldn't have those figures before the fight takes place. By the time these stats are known for a given fight, the outcome of the fight has already been determined, and therefore betting on the outcome becomes a moot point. In one of the other research papers that we reviewed, Johnson [1] used a logistic regression model for forecasting the results of MMA fights and used as predictors for this model several different metrics calculated from the raw fight stats. These metrics were intended to describe the fighters' behavior on a deeper level and included striking ratio, total takedown percentage, and ground activity per takedown. While for any single fight these metrics alone may not be sufficient for predicting a fight's outcome ahead of time (as they are only known once the fight is completed), taking averages of each of these metrics from each fighter's previous bouts and using those as predictors for our models could be a promising avenue for future experimentation.

It would also be interesting to test these same models on MMA fights from other federations aside from the UFC and see if the classification accuracy was consistent. If the performance was comparable in both cases, then that would lend more confidence as to the robustness of these models.

Final Conclusions

To conclude, the team will answer the following questions stated at the beginning of the document:

Primary Research Question: Can we build a model that will predict the outcome of UFC fights with more accuracy than the traditional betting odds?: Yes, after testing different models (Logistic Regression, SVM, Random Forest and Neural Networks) and choosing a voting classifier, it was confirmed that we can predict UFC fights with a high accuracy of 82%, which is better than the models reviewed in other research papers.

Supporting question #1: Which features impact fight results the most? Based on Logistic Regression and Random Forest models used for feature selection, we found that historical wins, fight statistics, reach, weight, and height were some of the features with the highest significance to predict UFC fights outcomes.

Supporting question #2: How accurately can we predict the outcome? Based on the voting classifier where we took into consideration the different models tested models (Logistic Regression, SVM, Random Forest and Neural Networks), we can predict UFC fights with an 82% accuracy.

Supporting question #3: Are there any outliers in the data? How could/should this information be used to help coach fighters or teams? There were no outliers in the data, but there were missing values. Because there were very few rows with missing values, the team removed these rows from the dataset. Fight statistics such as stance and reach were standardized based on the length of the round for better consistency among the statistics. The UFC fights model information can help coach fighters because it can help them focus on the variables that proved to be significant for winning fights. For example, stance, reach, and fighter variables proved to be significant, therefore coaches could focus their training on these.

Works Cited

Johnson [1] used a logistic regression model for forecasting the results of MMA fights and produced results significantly better than baseline models. This model used as predictors simple "count" variables (e.g. number of takedowns, number of strikes, etc.), as well as other metrics calculated using these variables that aimed to describe fighter behavior on a deeper level. These calculated predictors included striking ratio, total takedown percentage, and ground activity per takedown. One advantage of this approach is that it allows for the outcome of a fight to be predicted before the fight takes place. It's one thing to predict the outcome of a fight given the total number of strikes landed by each fighter and other types of data along those lines, but obviously those numbers can't be known until the fight has already happened.

Sharma et al [2] used a multi-layer neural network model for predicting the outcomes of MMA fights using UFC data from 1993-2019, achieving a 10% increase in accuracy over previous studies. 117 different predictor variables were used, among which were the age and weight for each competitor, average takedown percentage, number of title bouts, and number of ground strikes attempted and landed. The final model produced 71% accurate classification results, although the authors noted that the input data was highly unbalanced, so that could be an area of opportunity for producing better performing models. Furthermore, this group only tested one type of model and didn't provide much detail as to how they

trained that model and what the final hyperparameters were, so that would be another easy area for improvement.

[3] This citation is from an open-source project published on github in 2020. The goal of this project is very similar to ours: predicting the outcome of UFC fights. After cleaning, pre-processing, and visualizing the data, this person used an ensemble method to test various types of models and combined them for the best results. They started with a deep neural network, and then tried other methods like KSVM, KNN, decision trees, Log Regression, and XGboost. I think this project could aid us in many ways. The first will see what their input data looked like and how they transformed it. Since we are starting with 6 separate csv files, this will be useful in determining how to properly transform ours. We can also benefit from seeing how they built their various models, and which features they ultimately ended up using. Unfortunately, there's no final write-up on the results or accuracy of this program. However, seeing all their input data and feature extractions processes will make this a useful source for us.

McQuaide [4], a Stanford undergraduate student, built a UFC fight prediction model and published their findings in this paper. They started by splitting the data using k-folds cross-validation. Then they tested multiple classification models such as Decision Tree, DNN, and gradient boosting. As it turns out, gradient boosting was performed with the highest accuracy on the test data, although the results were marginal. Their highest overall accuracy on the test data was 61.5%, with gradient boosting. A decision tree came close behind with 60.1%, which is solid, but not terribly convincing since it's only slightly better than chance. These results will be beneficial for us as we look to gauge the accuracy of our own models.

Johnson, J. D. (2012). *Predicting outcomes of mixed martial arts fights with novel fight variables* (thesis). https://getd.libs.uga.edu/pdfs/johnson_jeremiah_d_201208_ms.pdf

1. G. Sharma and A. K. Uttam, "Multilayer Neural Network Model for Mixed martial arts Winner Prediction," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702452. https://ieeexplore.ieee.org/abstract/document/9702452
2. Rezan-21. *UFC-Prediction (2020)* https://github.com/rezan21/UFC-Prediction
3. McKinley McQuaide (2019). *Applying Machine Learning Algorithms to Predict UFC Fight Outcomes* https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647731.pdf