Gary Gau Ku

**UFC Fight Analysis**

**Problem Statement**

What does it take to be a champion in the world of mixed martial arts? In the sport of professional fighting, what do you have to do to win? What traits and habits should you work on that will help improve your chances of winning? What actions should you take in a fight to maximize your chances of coming out on top?

Using a dataset found on Kaggle composed of fight data describing over 5,000 fights in the UFC we will use statistical techniques in an attempt to find what factors are key to attaining victory in the sports of mixed martial arts.

(Link to the data set: https://www.kaggle.com/rajeevw/ufcdata?select=raw\_total\_fight\_data.csv)

**Data Cleaning**

The original data set was taken from Kaggle and contained 5144 rows with 151 columns. After dropping rows with missing data as well as rows where the result of the match was a tie the number of rows was reduced to 3,648 which was still an acceptable amount of data to work with. When analyzing the rows with missing data, most of them seemed to come from untelevised matched where there was no way to collect data from them.

The data was recorded in a ‘red vs blue’ format where the columns would describe actions taken by either the red fighter or the blue fighter with the variable of interest describing whether the blue fighter or the red fighter wins. In the interest to dimension reduction as well as making the data easier to analyze, features were extracted from the existing features by taking the difference between the actions of the blue and red fighters. Also, the outcome of the match was one ’hot’ encoded into a Bernoulli variable describing whether or not the red fighter won. The resulting number of columns was 17 (with 16 independent variables and 1 dependent variable).

Additionally, the data set seemed to be imbalanced where the number of times the red fighter won a fight was disproportionally low compared to the times the red fighter lost. To address this issue the SMOTE method was used to oversample observations where red wins. The resulting data frame had 4,046 rows and 17 columns.

The variables of the final data frame are as follows:

* **diff\_age**: the difference in age between the 2 fighters
* **diff\_heigh**t: the difference in height between the 2 fighters
* **diff\_reach**: the difference in reach between the 2 fighters
* **diff\_avg\_body\_att**: the difference in the average number of body strikes attempted per round between the 2 fighters
* **diff\_avg\_body\_landed**: the difference in the average number of body strikes landed per round between the 2 fighters
* **diff\_avg\_clinch\_att**: the difference in the average number of clinches attempted per round between the 2 fighters
* **diff\_avg\_clinch\_landed**: the difference in the average number of clinches landed per round between the 2 fighters
* **diff\_avg\_distance\_att**: the difference in the average number of attempts to close distance per round between the 2 fighters
* **diff\_avg\_distance\_att**: the difference in the average number of successful attempts to close distance per round between the 2 fighters
* **diff\_avg\_ground\_att**: the difference in the average number of ground strikes attempted per round between the 2 fighters
* **diff\_avg\_ground\_landed**: the difference in the average number of ground strikes landed per round between the 2 fighters
* **diff\_avg\_head\_att**: the difference in the average number of head strikes attempted per round between the 2 fighters
* **diff\_avg\_head\_landed**: the difference in the average number of head strikes landed per round between the 2 fighters
* **diff\_avg\_kd**: the difference in the average number of knockdowns per round between the 2 fighters
* **diff\_avg\_leg\_att**: the difference in the average number of leg strikes attempted per round between the 2 fighters
* **diff\_avg\_leg\_landed**: the difference in the average number of leg strikes landed per round between the 2 fighters
* **diff\_avg\_pass**: the difference in the average number of guard passes per round between the 2 fighters
* **diff\_avg\_rev**: the difference in the average number of reversals per round between the 2 fighters
* **diff\_avg\_sig\_str\_att**: the difference in the average number of significant strikes attempted per round between the 2 fighters
* **diff\_avg\_sig\_str\_landed**: the difference in the average number of significant strikes landed per round between the 2 fighters
* **diff\_avg\_sig\_str\_pct**: the difference in the average number of significant strikes protected against per round between the 2 fighters
* **diff\_avg\_sub\_att**: the difference in the average number of submissions attempted per round between the 2 fighters
* **diff\_avg\_td\_att**: the difference in the average number of takedowns attempted per round between the 2 fighters
* **diff\_avg\_td\_landed**: the difference in the average number of takedowns landed per round between the 2 fighters
* **diff\_avg\_td\_pct**: the difference in the average number of takedowns protected against per round between the 2 fighters
* **diff\_avg\_total\_str\_att**: the difference in the average number of total strikes attempted per round between the 2 fighters
* **diff\_avg\_total\_str\_landed**: the difference in the average number of total strikes landed per round between the 2 fighters

**Exploratory Data Analysis**

After the data was cleaned and prepared for analysis, 4 predictive algorithms were tested during data exploration: kNN, logistic regression, decision tree, and random forest. The hyperparameters for each algorithm were optimized with a 5-fold cross validation grid search and a confusion matrix for each algorithm was constructed based on the results of applying the respective algorithm on a holdout set of 810 samples. The results are as follows:

**kNN Classifier**

Best Score: 0.7039210120709914

Best Parameters: {'n\_neighbors': 1}

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Red Win | 1.00 | 1.00 | 1.00 | 415 |
| Red Loss | 1.00 | 1.00 | 1.00 | 395 |

|  |  |  |
| --- | --- | --- |
| kNN Cf Matrix | Predicted red fighter win | Predicted red fighter loss |
| Actual red fighter win | 415 | 0 |
| Actual red fighter loss | 0 | 395 |

**Logistic Regression**

Best Score: 0.5991090967357964

Best Parameters: {'C': 10}

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Red Win | .63 | .64 | .63 | 415 |
| Red Loss | .62 | .61 | .61 | 395 |

|  |  |  |
| --- | --- | --- |
| Log Reg Cf Matrix | Predicted red fighter win | Predicted red fighter loss |
| Actual red fighter win | 265 | 150 |
| Actual red fighter loss | 155 | 240 |

**Decision Tree**

Best Score:0.6097416411054648

Best Parameters: {'criterion': 'entropy', 'max\_depth': 6}

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Red Win | .65 | .90 | .76 | 415 |
| Red Loss | .83 | .50 | .62 | 395 |

|  |  |  |
| --- | --- | --- |
| Tree Cf Matrix | Predicted red fighter win | Predicted red fighter loss |
| Actual red fighter win | 375 | 40 |
| Actual red fighter loss | 198 | 197 |

**Random Forest**

Best Score:0.6698136702833859

Best Parameters: {'bootstrap': False, 'max\_depth': 9, 'min\_samples\_leaf': 3, 'min\_samples\_split': 3}

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Red Win | .95 | 1.00 | .98 | 415 |
| Red Loss | 1.00 | .95 | .97 | 395 |

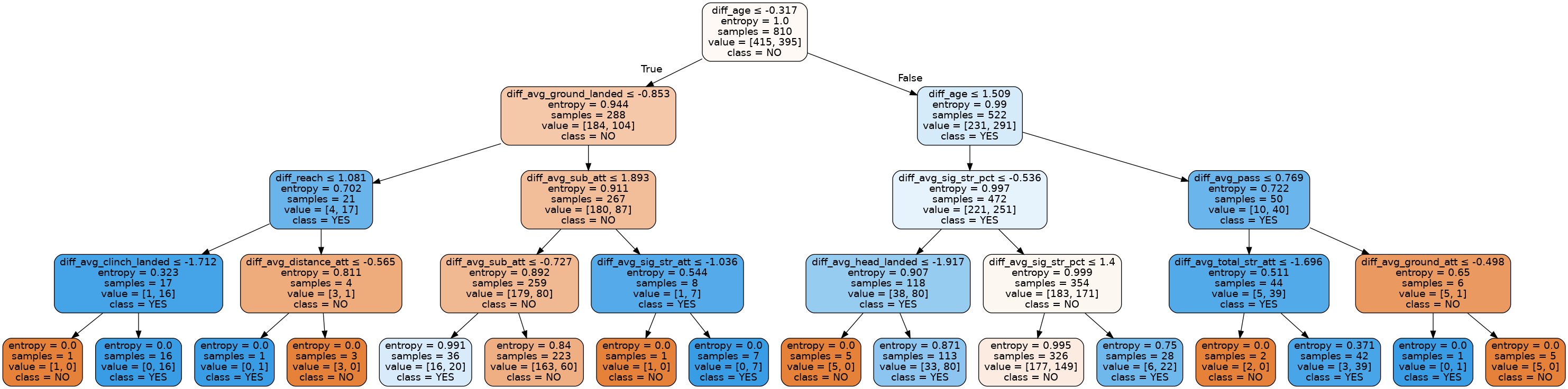
|  |  |  |
| --- | --- | --- |
| Ran Forest Cf Matrix | Predicted red fighter win | Predicted red fighter loss |
| Actual red fighter win | 415 | 0 |
| Actual red fighter loss | 21 | 374 |

**Model Selection**

Of the models tested the model that seemed to perform the best was kNN classification. However, at 100% accuracy, the results of the kNN algorithm seem dubious. This may be due to the fact that the data set was balanced using the SMOTE method which uses Euclidean distance as a metric for producing synthetic data points. Because the kNN algorithm also uses Euclidean distance to classify data, the use of SMOTE may have compromised the applicability of the kNN classifier. When testing the kNN classifier without using SMOTE, its accuracy was much lower. Thus, despite its seemingly high performance, the kNN model was not chosen for further analysis.

Logistic regression had the worst performance out of all the models. This may be attributed to the fact that the format of the data set was in a ‘red vs blue’ format. Because of the way the dependent variable was formatted, the resulting ‘red wins’ Bernoulli variable extracted from the ‘winner’ column was arbitrary in that a ‘1’ could be represented as a ‘0’ if the values for the independent variables were multiplied by -1 and vice versa. Because of this, the logistic regression model may not have been suited to this classification task.

The decision tree classifier performed reasonably with an overall accuracy of around 71% and thus was selected for further investigation. Based on the results of a decision tree with a max depth of 4, the difference between 2 fighters’ ages seems to be the most important factor as to whether they win or lose followed by the difference in the number of ground strikes landed, submissions attempted, significant strikes protected against, successful guard passes, and reach between the 2 fighters. These results are shown in the following decision tree diagram:

Figure 1: Decision Tree Diagram

Furthering the application of the decision tree classifier, a random forest algorithm was run producing promising results. With an overall accuracy of 97%, the random forest algorithm performed the best out of every algorithm tested. Additionally, the results of the random forest model seems reasonably consistent with the results of the decision tree classifier highlighting the importance of age in determining the results of a fight. This result is shown in the following variable importance plot derived from the random forest model:

Figure *2*: Variable Importance Plot

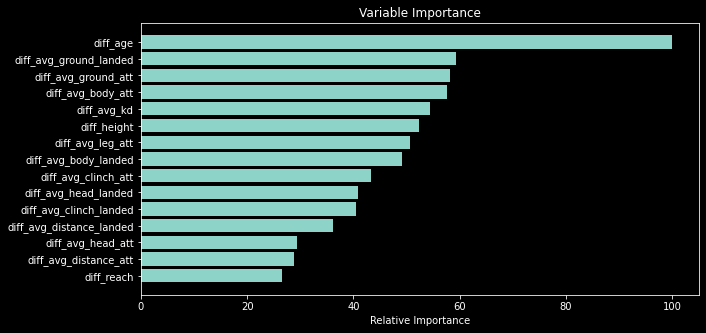
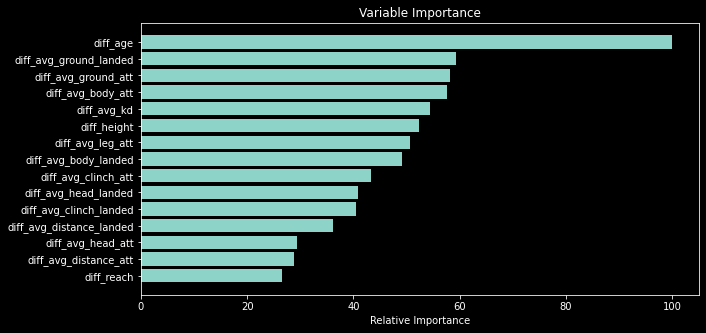


Figure 2: Variable Importance Plot



Additionally, logistic regression plots were constructed for the top 3 variables listed on the variable importance plot:

Figure 3: diff\_age vs. Red Win

Figure *2*: Red Wins vs. diff\_age

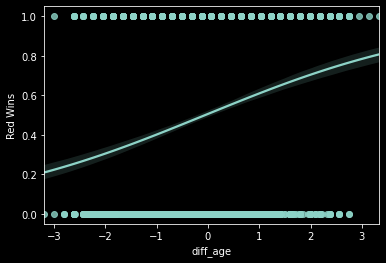
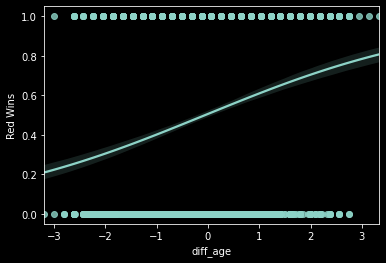


Figure *3*: Red Wins vs. diff\_ground\_landed

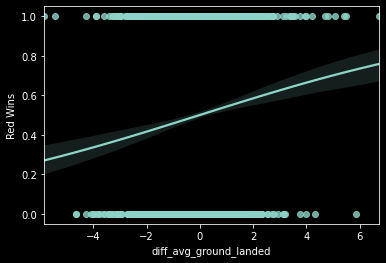
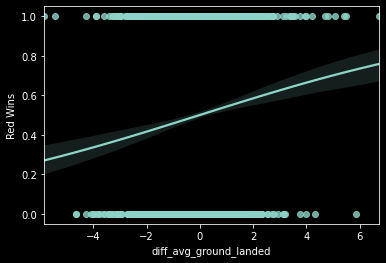
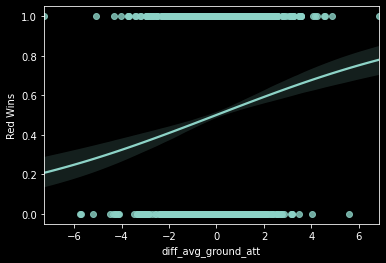
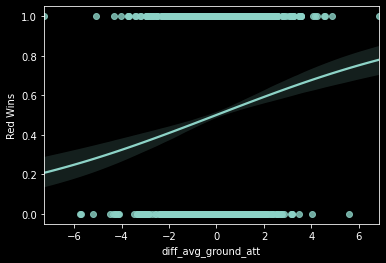


Figure 4: Red Wins vs. diff\_ground\_att



Out of these 3 variables, the difference in age between the 2 fighters seems to be the significant variable that can reasonably be used to predict the outcome of the fight. It appears that the older the fighter is relative to their opponent the more likely they are to win a fight. There does also seem to be a trend where the fighter who attempts and land more ground strikes is more likely to win the fight.

**Conclusion**

Out of the algorithms tested, the algorithm that seemed best suited to predicting the outcome of fights given the difference in actions taken between the fighters. Additionally, it seems that fighters that are older than their opponents tend to be more likely to win suggesting that experience tends to triumph over youth. It may also suggest that younger fighters tend to be the ones that are ‘weeded out’ leading to better records for older fighters. Analyses of actions within a fight seem to suggest that many fights are won on the ground as fighters who strike grounded opponents more often tend to have a greater chance of winning.

For future research, it may be worth trying more algorithms and using different ways to balance the data set. Also, feature selection algorithms could be used for dimension reduction rather than deriving features from the existing feature. Using the original features could shed some more insight on what it takes to win a fight.