UFC Fight Finish Prediction

Classification Project Report

Data Science T5 Boot camp 2021

Name: Faisal Mirza

Gmail: faisal.a.mirza11@gmail.com

GitHub: faisalmirza11

Abstract:

In this classification project, I used UFC fights data from Kaggle to predict a fights finish i.e., knockout, decision, or submission by seeing data of a match, like number of punches, number of takedowns, number of rounds, total fight time, fighter average strikes per minute and more.

Framing question: Will the fight finish by knockout, or by decision, or by submission?

Design:

UFC is a company that organizes mix martial arts fights (MMA), and it is when basically two fighters get in an octagon cage and fight, the fight will end in three ways, either by knockout or by decision or by submission.

Data:

The dataset has been obtained from Kaggle, and it contains 4885 rows and 119 columns. Some of these columns include:

- Red corner fighter name, Blue corner fighter name, fight date, location, weight class
- Number of rounds
- B and R average significant strikes landed, average significant striking accuracy, Average Submissions Attempted, Average takedowns landed, Takedown accuracy
- Fighter height, reach and weight, age
- The finish of the fight i.e., knockout, unanimous decision, or submission
- Fight finish round

Scope:

I dropped about 250 rows and used 4631 rows, I used 44 columns that I felt like they were going to add value to the model.

Algorithms:

I started by viewing the dataset to understand it clearly, then after looking at the data and understanding it, I started to look for issues and problems in the data that required cleaning, and some of the cleaning that I did includes:

- Dropped 16 rows where target column was: **DQ, Overturned**
- Dropped 240 rows where target (finish) was **null**

I made a subset (df2) of only the columns that I thought were going to be useful to the models.

Feature engineering:

- Converting **finish round time** column to integer
- Replacing values of 0 with the column mean
- There are three types of decision finish U-DEC, S-DEC, M-DEC, I made them all as DEC

I split the data to 80% training, 20% validation, then after cleaning the dataset and performing the feature engineering I started to experiment using Classification models to find the best score, and some of the models I used include:

- Baseline logistic regression model
- Logistic regression down sample
- KNN
- Decision Tree
- Random Forest
  In the end the best model was Random Forest

Tools:

Technologies: Python, Jupiter notebook

Libraries: Sklearn, statsmodel, imblearn.over_sampling, imblearn.over_sampling, Numpy, Pandas, Matplotlib, Seaborn.

Communication:

Everything related to this project will be available on my GitHub at faisalmirza11 in the repository DS-T5-projects/Classification project.

Baseline Logistic Regression Model:

Training scores:
accuracy score: 0.75
f1 score: 0.63
precision score: 0.66
recall score: 0.64

Validation scores:
accuracy score: 0.78
f1 score: 0.67
precision score: 0.71
recall score: 0.67

```
Logistic Regression confusion matrix:

 [[244  22  23]
 [ 21 447  10]
 [106  14  40]]
```