IST707

PREDICTING UFC FIGHT OUTCOMES USING MACHINE LEARNING

Laxman Kumar
lakumar@syr.edu
School of Information Studies
Syracuse University

Saheb Singh
ssingh55@syr.edu
School of Information Studies
Syracuse University

Abhiraj Singh
asingh73@syr.edu
School of Information Studies
Syracuse University

Introduction

Sports betting is a $155 billion industry. Fighting ranks among the top in the industry, and the Ultimate Fighting Championship (UFC) is currently taking steps to push it even further. Mixed Martial Arts (MMA) fighter statistics involve everything from skill centric values such as wins, and significant strikes landed to physiological measurements such as height and reach. There are over one hundred different features up to analyze before any given fight, and machine learning can be used to best understand which are most relevant, and to indent trends and predict the outcomes (win/draw/loss) of each fight.

Problem Statement

The goal of this study is to explore our ability to predict the outcome of UFC fights based on each match's pre-fight statistics using machine learning models. An accurate prediction model could both inform the best placed bets (and potential risk associated) for each fight, but also could provide insight to coaches when accepting fights to begin with, simply by looking at the opponent's statistics relative to their fighter. It could also be used to help to identify which features are most significant in this prediction.

Dataset and Features

There are over one hundred different fighter statistics on UFC Stats for each of the fights in UFC record from 1993 to 2019, which include information such as fighters' height, weight, reach, and stance, as well as statistics such as win streaks, strike percentage, guard passes, and strikes landed by location. The dataset used in this study was scraped from UFCStats and statistics pertaining to the circumstances of the fight (i.e. location, number of allocated rounds, etc.) were removed to as they were deemed irrelevant and did not align with the fighter-centric intention of the study.

Dataset link : https://www.kaggle.com/rajeevw/ufcdata

Column name	Description
R *and B*	prefix signifies red and blue corner fighter stats respectively
opp	containing columns is the average of damage done by the opponent on the fighter
KD	is number of knockdowns
SIG_STR	is no. of significant strikes 'landed of attempted'
SIG_STR_pct	is significant strikes percentage
TOTAL_STR	is total strikes 'landed of attempted'
TD	is no. of takedowns
TD_pct	is takedown percentages
SUB_ATT	is no. of submission attempts
PASS	is no. times the guard was passed?
REV	*Probably reversels*
HEAD	is no. of significant strinks to the head 'landed of attempted'
BODY	is no. of significant strikes to the body 'landed of attempted'
CLINCH	is no. of significant strikes in the clinch 'landed of attempted'
GROUND	is no. of significant strikes on the ground 'landed of attempted'
win_by	is method of win
last_round	is last round of the fight (ex. if it was a KO in 1st, then this will be 1)
last_round_time	is when the fight ended in the last round
Format	is the format of the fight (3 rounds, 5 rounds etc.)
Referee	is the name of the Ref
date	is the date of the fight
location	is the location in which the event took place
Fight_type	is which weight class and whether it's a title bout or not
Winner	is the winner of the fight
Stance	is the stance of the fighter (orthodox, southpaw, etc.)
Height_cms	is the height in centimeter
Reach_cms	is the reach of the fighter (arm span) in centimeter
Weight_lbs	is the weight of the fighter in pounds (lbs)
age	is the age of the fighter
title_bout	Boolean value of whether it is title fight or not
weight_class	is which weight class the fight is in (Bantamweight, heavyweight, Women's flyweight, etc.)
no_of_rounds	is the number of rounds the fight was scheduled for

Column name	Description
current_lose_streak	is the count of current concurrent losses of the fighter
current_win_streak	is the count of current concurrent wins of the fighter
draw	is the number of draws in the fighter's ufc career
wins	is the number of wins in the fighter's ufc career
losses	is the number of losses in the fighter's ufc career
total_rounds_fought	is the average of total rounds fought by the fighter
total_time_fought(seconds)	is the count of total time spent fighting in seconds
total_title_bouts	is the total number of title bouts taken part in by the fighter
win_by_Decision_Majority	is the number of wins by majority judges decision in the fighter's ufc career
win_by_Decision_Split	is the number of wins by split judges decision in the fighter's ufc career
win_by_Decision_Unanimous	is the number of wins by unanimous judges decision in the fighter's ufc career
win_by_KO/TKO	is the number of wins by knockout in the fighter's ufc career
win_by_Submission	is the number of wins by submission in the fighter's ufc career
win_by_TKO_Doctor_Stoppage	is the number of wins by doctor stoppage in the fighter's ufc career
avg	average over number of rounds in the fight

Approach

We have followed the Crisp DM – Cross industry standard procedure for data mining methodology. This is the most widely used analytics model. This model has five major phases –

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation

Modeling & Evaluation

We are trying to predict outcome of an UFC match using fighter's pre-fight statistics. For our problem statement we are planning to implement Regression, Random Forest and XGBoost. For

evaluation of these models we will be using the following metrics such as ‘Recall’, ‘F-1 Score’, ‘Precision’, and ‘Accuracy’. Our main focus will be to improve the ‘F-1 score’ of our model as we want to lower the number of ‘false-negative’ and ‘false-positive’ so that user is informed and places the best bets. It is necessary for our model to be able to predict correctly the result of the fight, predicting an incorrect fighter as the winner defeats the overall purpose of our model as user then won’t be able to use our model for betting.

Reference

1. https://www.kaggle.com/rajeevw/ufcdata