UFC Capstone

Understanding what matters and predicting fight outcomes

By Bing Chen

History of the UFC and MMA

- Started in 1993 as a tournament to see which fight style was the "best"
- In reality, it was created to showcase "Gracie Jiu Jitsu" which relied on chokes and joint locks to force the opponent to "submit" or give up.
- No weight classes, no rules, no time limits
- Royce Gracie used jiu jitsu to defeat much larger and stronger opponents

Now:

- UFC is a sport with a commissioning body, strict rules.
- MMA (Mixed martial arts) is the fastest growing sport in the world

Understanding the basic rules

- Two fighters at a time in the cage
- Each round of fighting is 5 minutes
- Normal fights last 3 rounds. Championship fights last 5 rounds
- Fighters can win by knocking out their opponent or causing their opponent to submit or “tap out”
- If neither fighter is knocked out or submits, then the winner is declared by 3 judges who score each round

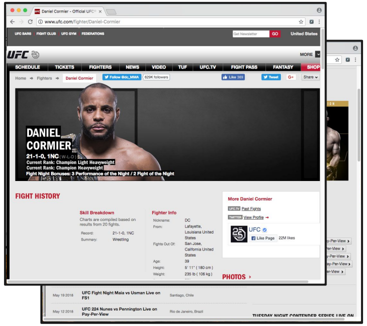
Objectives for the Project

- Create a model that can predict the winner of a fight between 2 UFC fighters
- Understand what aspects about a fighter or a matchup influences the probability of winning
- Explore UFC data and look for interesting insights
- Everyone has some “subject matter expertise” when it comes to fighting. We intuitively have a guess about what matters in a fight. In this project I would like to examine our intuitions and see what the data actually says.

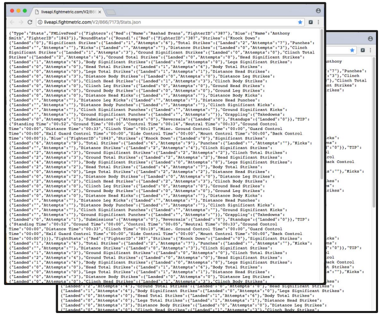
Collecting the Data

Collected data by scraping from 2 primary sources:

1. **Official UFC website**
    a. Events
    b. Fights
    c. Fighter Info



2. **FightMetrics.com API**
    a. “V1” Json - High level information for each fight
    b. “V2” Json - Detailed fight statistics



Understanding the Data

After scraping all the resources, the data we have really falls into 2 categories:

1. Static Fighter Information - Height, Weight, Reach, Age, Fighting Style
2. Historical Fight Statistics: (From V2 JSON)

The historical fight stats can be further broken up into a few categories:

1. Keys (EventId, FightId, FighterIds)
2. Striking Metrics
3. Grappling Metrics

Understanding the Data

Grappling Metrics - Takedowns, Standups, Submissions

Striking Metrics:

- **Target of Strike:** Head, Body or Legs
- **Position:** Distance, Clinch, Ground

attempted	distance	clinch	ground
head	78.0	9.0	26.0
body	29.0	5.0	3.0
leg	17.0	3.0	1.0

landed	distance	clinch	ground
head	23.0	3.0	18.0
body	20.0	5.0	2.0
leg	15.0	3.0	1.0

Preparing the Data: Key to the project

Feature Engineering and Data Preparation:

The key to the project is engineering features that are predictive of wins since the base set of scraped features have many issues:

- Only “attempted” and “landed”, not percent landed
- The fight stats are in absolute numbers for an entire fight
- The differential between fighters might be a big indicator
- The ratio of the types of strikes might matter
- The quality of opponent matters
- Knockdowns / Significant Head Strikes Landed ⇒ Measure of power
- Standups / opponent_take_downs => Measure of take down defense

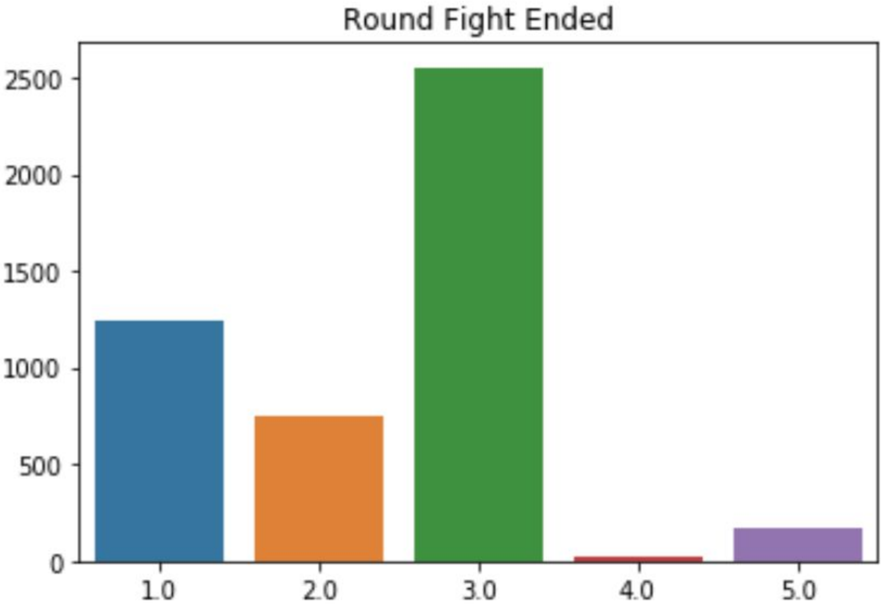
Preparing the Data: Time Series

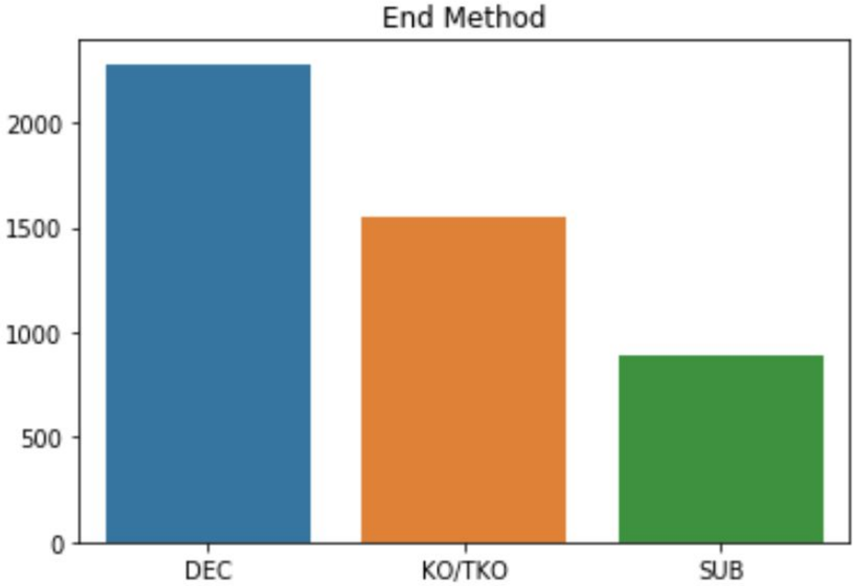
Feature Engineering and Data Preparation:

- The data as it was scraped, cannot be put into a predictive model because the fight stats are only known after the fight is over.
- In order to use the fight stats, we must only use the statistics leading up to a fight we want to predict.
- Transformation: Calculating the historical average for each fight

Visualizations

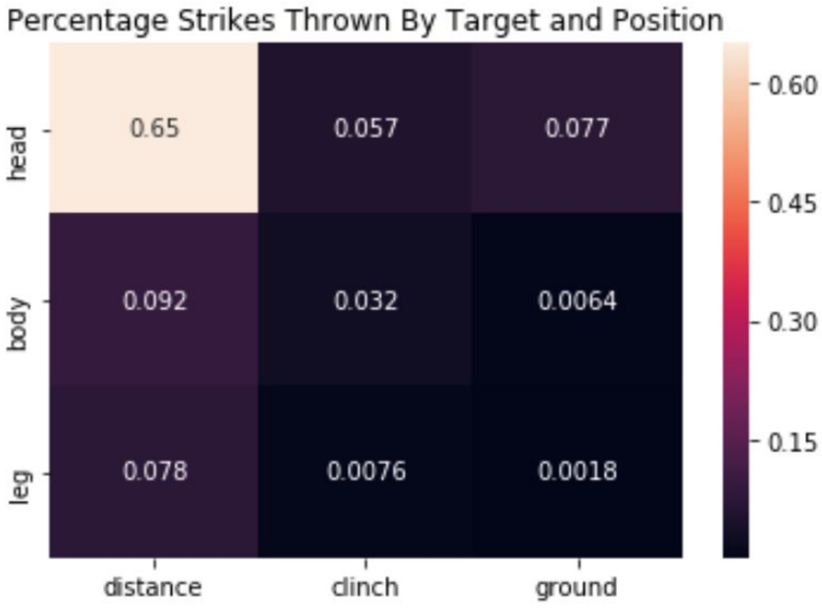
How and when the fight ended





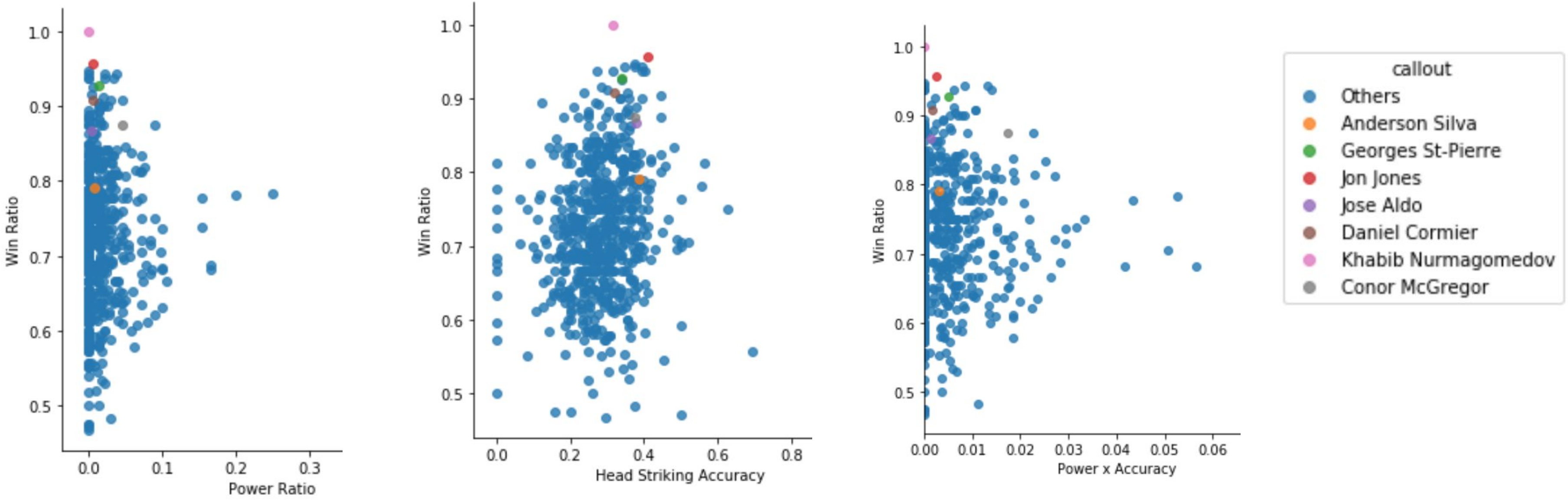
Visualizations

Fighters usually aim for the head from distance.



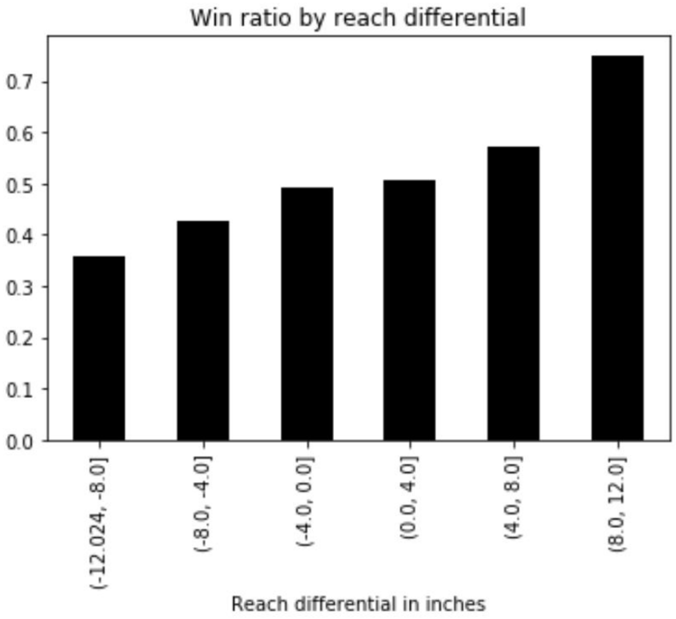
Visualizations

Is being good at striking to the head from distance a predictor of win rate?



Visualizations

Reach advantage vs. win ratio



Modeling: Logistic Regression

- Best model achieved a 58% accuracy rate
- Tried selecting top 100 features and principle component analysis
- Top 3 positive and negative predictors shown

f1_reach_adv	0.044867
f1_grappling_submissions_attempts_avg_diff	0.044572
f1_knock_down_landed_avg_diff	0.041376
f2_total_strikes_landed_avg_diff	-0.022935
f2_head_total_strikes_landed_avg_diff	-0.023089
f1_f2_clinch_head_strikes_landed_avg	-0.038542

Modeling: Random Forest

- Best model achieved a 60% accuracy rate
- Tried selecting top 100 features and principle component analysis
- Top 6 feature importance

f2_head_significant_strikes_landed_diff_avg	0.031720
f2_head_significant_strikes_percent_avg_diff	0.030295
f1_head_significant_strikes_landed_avg_diff	0.029473
f1_significant_strikes_landed_diff_avg	0.028200
f1_f2_clinch_head_strikes_percent_avg	0.024601
f1_significant_strikes_attempts_diff_avg	0.023133

Current Issues and Next Steps

- **The biggest challenge was the lack of enough quality training data**
    - Time in Position columns had to be dropped due to too many nulls
    - Actual data from V2 Jsons was missing about ⅓ of the fights for each fighter
    - 1st recorded fight for every fighter had to be dropped
    - Only ended up with about ~2600 rows of training data with ~700 features
- **Next steps**
    - Try training a model that doesn't use any in fight statistics
    - Try training a model that looks at who each fighter has defeated in order to predict
        - Page Rank Algorithm
    - Better feature engineering with the data I have
    - Build more exotic models and ensemble together
    - Compare to betting odds if I can find the data