Predicting Boxing Bouts

Using Machine Learning to attempt to predict professional boxing matches

Erik Euler

The Data and its Cleaning

- Initial data set had 387,427 entries across 26 columns
  - Limits had to be instituted for the fighter characteristic columns (Age, Height, Reach, etc.)
  - The data is scraped from a website that has boxing records going back to the late 1800s. Some of these fighter characteristics or statistics were not always collected
  - The next cleaning step was to remove the NA's. Instituting the limits also helped address the NA values because many of them were under fighter characteristic columns
  - There were also a few fight decisions that should not be considered in creating this mode:
    - RTD (Retired) - if a boxer retired mid-bout, this is the type of fluky decision that a model could not predict
    - DQ (Disqualification) - Also a decision that is fluky and one in which a model would struggle to predict
    - TD (Technical Decision) - fight is ended because of one fighters use of a headbutt, would group this decision with DQ, RTD
    - NWS (Newspaper Decision) – Out of data decision where the local newspaper would declare the winner of the fight (was often very biased)

decision	ageDiff	weightDiff	heightDiff	reachDiff	total_bouts_A	total_bouts_B
SD	8.0	0.0	4.0	-1.0	37	50
UD	-5.0	0.0	-10.0	-6.0	49	52
KO	-4.0	0.0	1.0	-1.0	47	34
SD	-8.0	0.0	0.0	-9.0	44	20
TKO	-6.0	1.0	-2.0	4.0	40	34

won_A	won_B	lost_A	lost_B	drawn_A	drawn_B	kos_A	kos_B	result
37	49	0	1	0	1	33	34.0	draw
48	50	1	2	1	1	34	32.0	win_A
46	31	1	3	1	0	32	19.0	win_A
43	19	1	1	1	2	31	12.0	win_A
40	30	0	4	1	0	29	18.0	win_A

AL M'COY IS TROUNCED BY GEORGE CHIP

New York, April 7—George Chip of Newcastle, Pa., last night outfought Al McCoy of Brooklyn in their ten-round battle at the Broadway A. C. in Brooklyn. Just about a year ago McCoy knocked out Chip in one round at the same club.

After a lot of dodging, McCoy agreed to box Chip again, and is sorry he did not hold on to his easily won reputation a little longer, for last night at the Broadway sporting club of Brooklyn, Chip gave his one-time vanquisher a lacing he will remember for some time to come. For ten rounds Chip punched McCoy all around the ring in the most approved manner, flooring him twice in the ninth round for the count.

That Chip did not return the compliment and knock out McCoy was due largely to overanxiety. The Pennsylvania boxer had McCoy staggering around the ring and practically all in, but he swung wild and used poor judgment in the last two rounds when the spectators looked for an ending of the bout before the time limit.

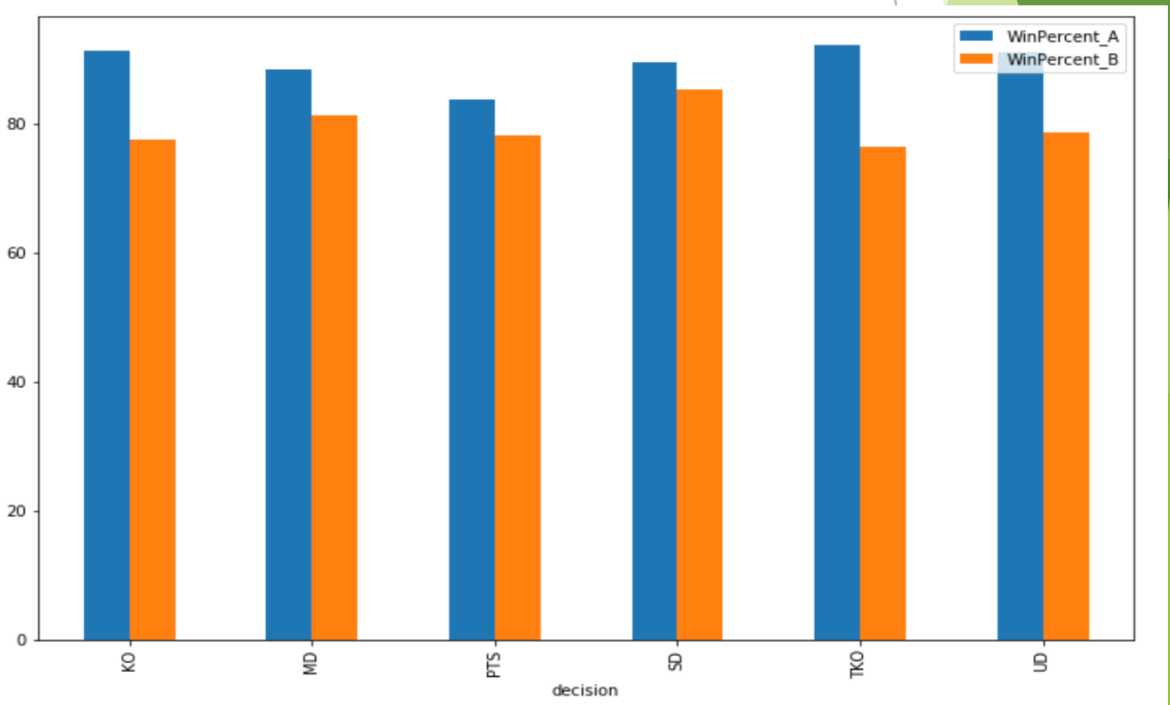
Final Cleaning

- Judges scores were also removed because of many incomplete observations
- Creation of “Difference” columns:
  - Instead of having this all this data for Boxer A, and B, I created “Diff” columns
  - Boxer A – Boxer B too create: AgeDiff, WeightDiff, ReachDiff, and HeightDiff
- Also created Total_Bouts, Win_Percent, and KO_Percent
- Cleaned Data: 6,643 rows, 16 columns

result	decision	ageDiff	weightDiff	heightDiff	reachDiff	total_bouts_A	total_bouts_B	WinPercent_A	WinPercent_B	KoPercent_A	KoPercent_B
draw	SD	8.0	0.0	4.0	-1.0	37	50	100.000000	98.000000	89.189189	68.000000
win_A	UD	-5.0	0.0	-10.0	-6.0	49	52	97.959184	96.153846	69.387755	61.538462
win_A	KO	-4.0	0.0	1.0	-1.0	47	34	97.872340	91.176471	68.085106	55.882353
win_A	SD	-8.0	0.0	0.0	-9.0	44	20	97.727273	95.000000	70.454545	60.000000
win_A	TKO	-6.0	1.0	-2.0	4.0	40	34	100.000000	88.235294	72.500000	52.941176
win_A	UD	-19.0	0.0	1.0	-1.0	39	53	100.000000	86.792453	74.358974	73.584906
win_A	TKO	-11.0	0.0	-5.0	-9.0	38	37	100.000000	89.189189	73.684211	75.675676
win_A	TKO	-13.0	1.0	8.0	6.0	31	32	100.000000	96.875000	74.193548	71.875000
win_A	TKO	-8.0	0.0	-5.0	-9.0	25	25	100.000000	92.000000	72.000000	84.000000
win_A	TKO	-10.0	0.0	2.0	-4.0	22	17	100.000000	94.117647	68.181818	88.235294

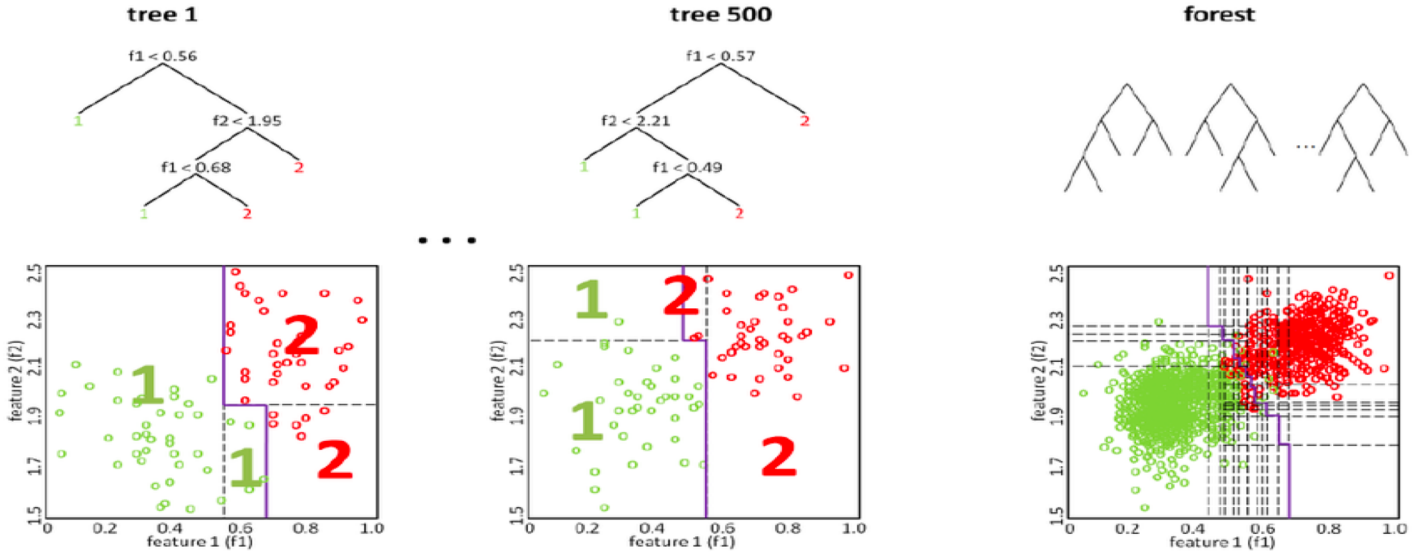
Initial Observations

- Initial Observations:
  - On average fighter A has a better win percentage for all decisions except for SD (split decision). This tells me that fights that were a split decision tend to be evenly matched fights
  - The mean Experience (total_bouts) seems to be much higher for fights decided by PTS for both fighter A and B, rest of the decision types are fairly normal in terms of Experience (total_bouts)
  - For TKO and KO decision, fighter A on average wins a fight with a KO more than 63% of the time
  - KO percentage is the least in fights decided by PTS, for both fighter A and fighter B
  - There is a fairly noteworthy difference in win percentage from fighter A to fighter B in fights decided by KO, TKO, and UD.
    - TKO, difference is 15.73%, KO, difference is 13.72%, and UD, difference is 12.58%.
  - Fight winners on average are younger than fight loser. By more than 2 years for a boxer A win, and by more than 1.25 years for a boxer B win



Model Creation – Predicting Result

- The most accurate model I created for predicting result was a ***Random Forest Classifier***
- Preformed a Train/Test split on the cleaned data set
- The resulting model accurately predicted the result in the testing data **84.82%** of the time
- Potential outcomes: Win_A, Win_B, and Draw
- Random Forests work in that they subset the data, then makes decisions on the various subsets
- The subsets are then aggregated
- Random Forests work well because by aggregating the many subsets and their decision trees, it will compensate for the potential subset might have been exposed to an unpredictable/extraneous noise factor



Model Creation – Predicting Decision

- The most accurate model for predicting decision was also a ***Random Forest Classifier***
- The data was split in the same fashion as for predicting the result, but this time the potential outcomes are:
  - *SD - Splitted Decision*
  - *MD - Majority Decision*
  - *UD - Unanimous Decision*
  - *KO - Knock Out*
  - *TKO - Technical Knock Out*
- The resulting Random Forest model accurately predicted the result in the testing data **50.69%** of the time
- Baseline Accuracy: **34.51%**
- ***Features Used:***
  - *Predicting Result: Age Difference, Weight Difference, Height Difference, Reach Difference, Total Bouts - Fighter A, Total Bouts - Fighter B, KO % - Fighter A, KO% - Fighter B, Win% - Fighter A, Win% - Fighter B, Draw % - Fighter A, and Draw % - Fighter B*
  - *Predicting Decision: Age Difference, Weight Difference, Height Difference, Reach Difference, Total Bouts - Fighter A, Total Bouts - Fighter B, KO % - Fighter A, KO% - Fighter B, Win% - Fighter A, Win% - Fighter B, Draw % - Fighter A, Draw % - Fighter B, and Result*

Conclusion & Next Steps

- Data incompleteness, my cleaning process reduced the amount of data by a lot more than I expected
  - I was also not expecting for the data to be skewed toward “result = Win_A”
- Additional Data: Punches Landed, an elusiveness data point, world ranking at the time of fight for both boxers, venue advantage (hometown of fighter)
- Analytics in sports is a fast-growing industry and I think there is an opportunity for advanced data collection and analysis in the sport of boxing
- Next Steps:
  - Attempt other classification models: Neural Networks, Support Vector Machines, etc.
  - Append this data with more fight data that’s out there too add more predictive variables