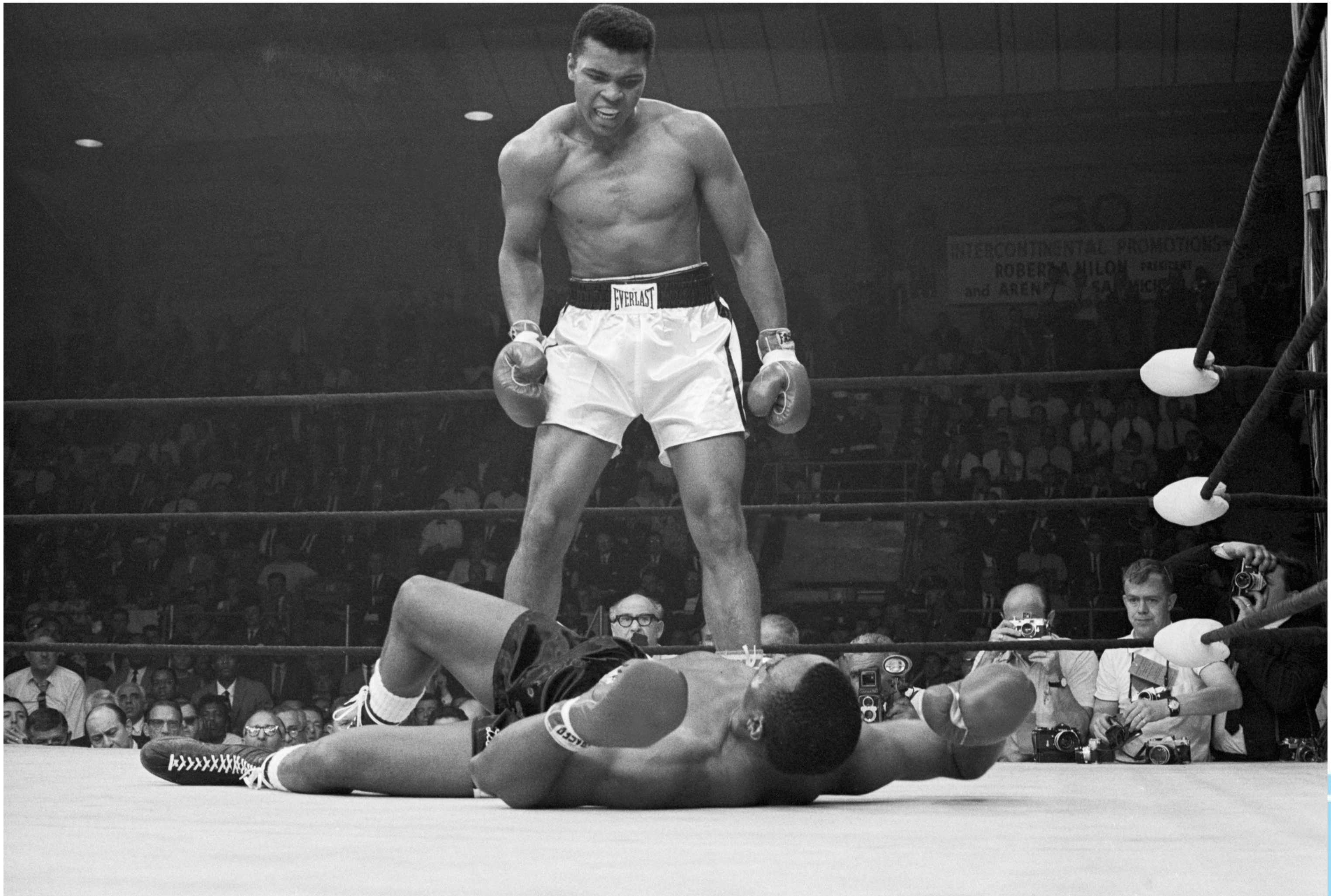
Predicting Boxing Winners



The Problem

- How can we predict the winner of a boxing match?

Who is interested?

- Sports analysts/broadcasters
  - More accurate predictions means higher ratings
- Sports betting facilitators
  - Better formulation of betting odds
- Boxing managers, trainers
  - Information about the opposing boxer might motivate certain training styles

The Data

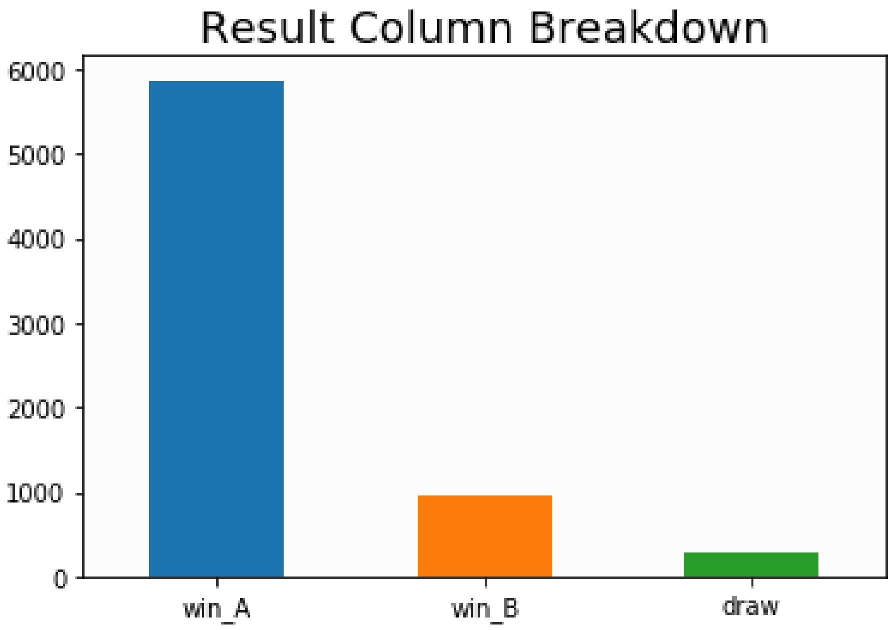
- Comma-separated file (CSV) of information about various boxing matches:
  - Physical attributes of each boxer
    - Age, height, weight, reach, stance
  - Win-loss records of each boxer
  - Information about the result of the match (how match ended, judges’ scoring, etc.)

Data wrangling/cleaning

1) Keep only rows where all physical attribute columns have valid entries
2) For each numerical column:
   1) Fill any values outside of the 1^{st} and 99^{th} percentile with the median of the column
3) Remove unneeded columns:
   1) Stance, since all matches were between boxers of the same stance
   2) Judges' scoring, since we are only interested in who won the match

Further data cleaning

- Remove all matches where the result was “draw”
  - These matches make up only 4% of the data
- Going forward in our analysis, we will focus on Boxer A, since this boxer won most of the matches

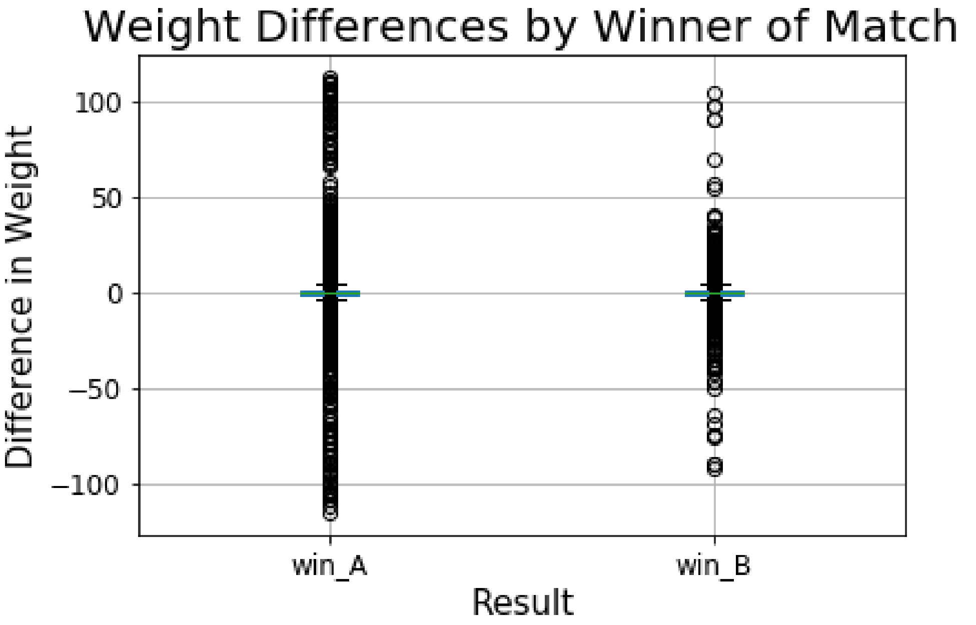


Four new columns

- We are interested in how each boxer measures up compared to his opponent.
- Creation of four new numerical columns for the differences in age, height, weight and reach:
  - diff_weight
  - diff_height
  - diff_reach
  - diff_age

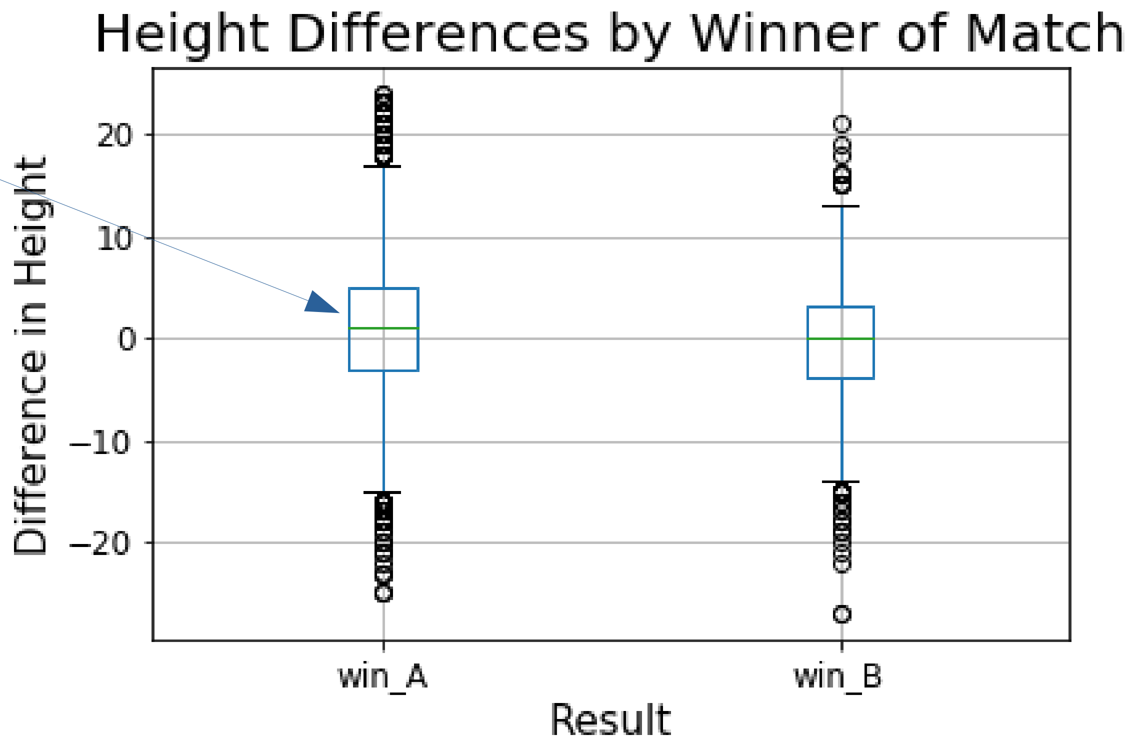
diff_weight EDA

- No apparent difference.
- Boxers tend to fight at the higher extreme of their weight class, so we did not expect a significant difference in these distributions.



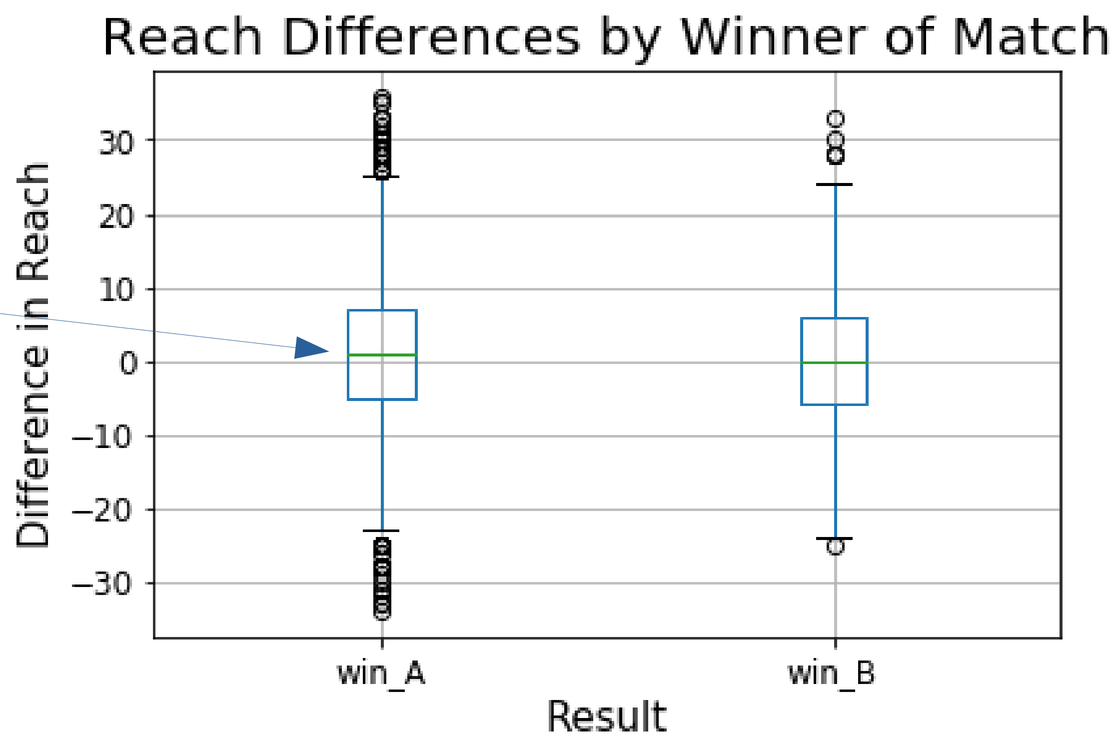
diff_height EDA

- When Boxer A wins, more than 50% of the time, he was taller than his opponent (positive difference in the heights).



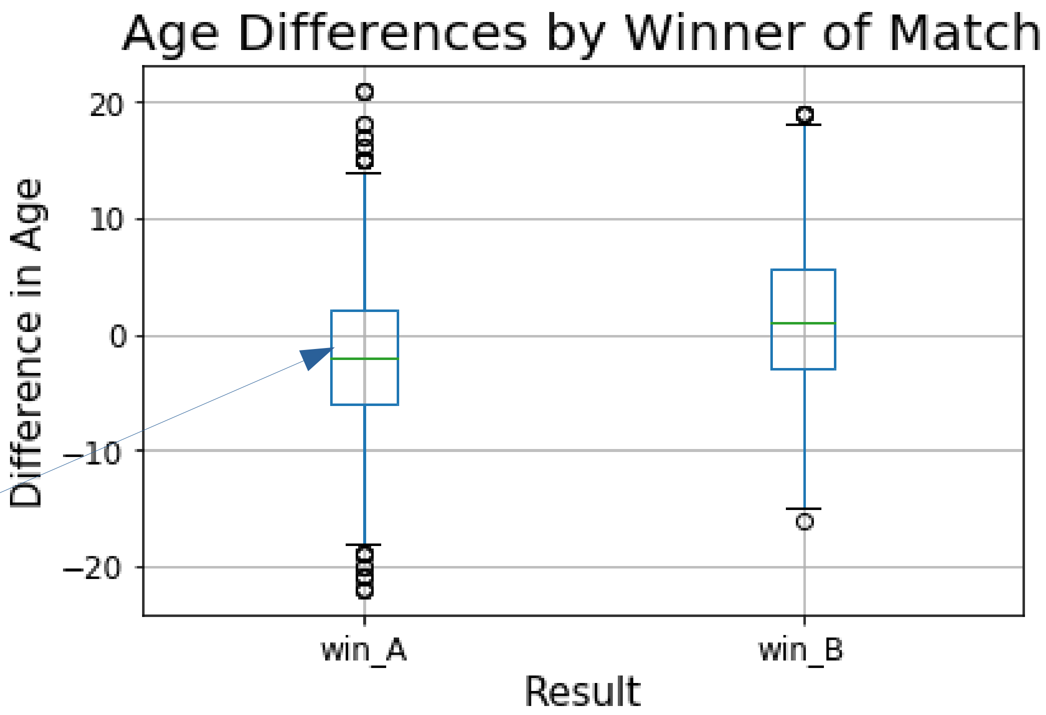
diff_reach EDA

- When Boxer A wins, more than 50% of the time, he was longer than his opponent.



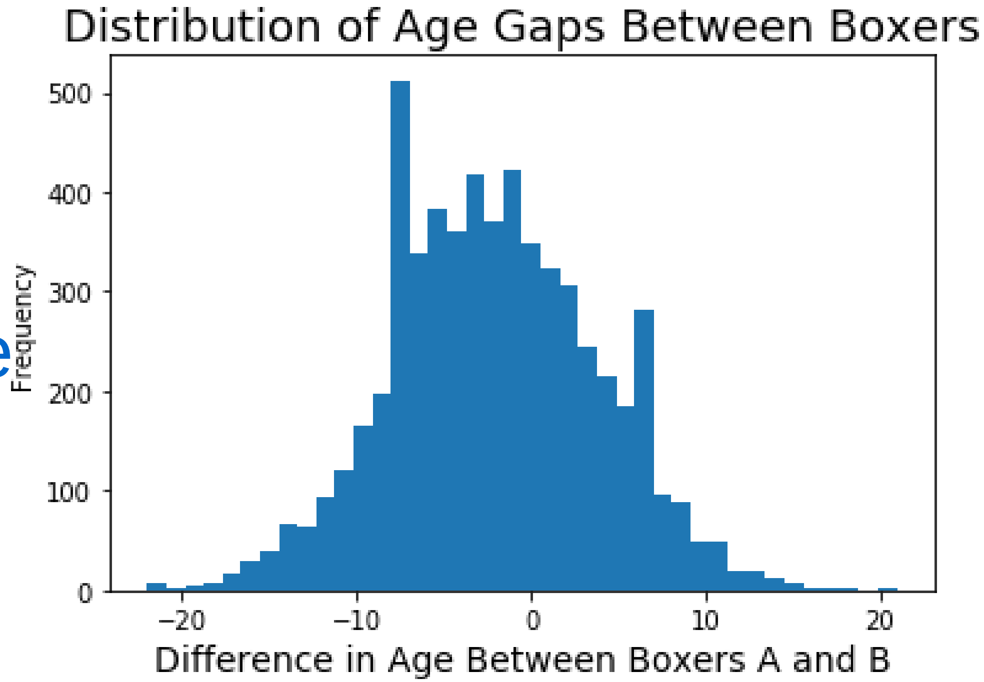
diff_age EDA

- Very significant difference in two distributions!
- Nearly 75% of the time Boxer A won, he was younger (negative difference) than his opponent.



diff_age EDA (cont'd)

- This histogram shows the age gaps for matches where Boxer A won.
- We see that the vast majority of matches were won when Boxer A was between 1 and 10 years younger than his opponent



Are the differences significant?

- Tested using a permutation hypothesis test for differences in the two distributions.
  - H_0: There is no difference in the two distributions.
  - H_A: There is a significant difference in the two distributions
- Very small p-values denote statistically significant difference in distributions.

Are the differences significant?

- Summary of hypothesis tests:
  - diff_weight → NOT statistically significant
    - Given the boxplots for the differences in weight, we are not surprised that this test was insignificant.
    - Since the differences in weight when Boxer A wins are no different from when Boxer B wins, we will be excluding this column from our predictive features.
  - diff_height → statistically significant
  - diff_reach → statistically significant
  - diff_age → statistically significant

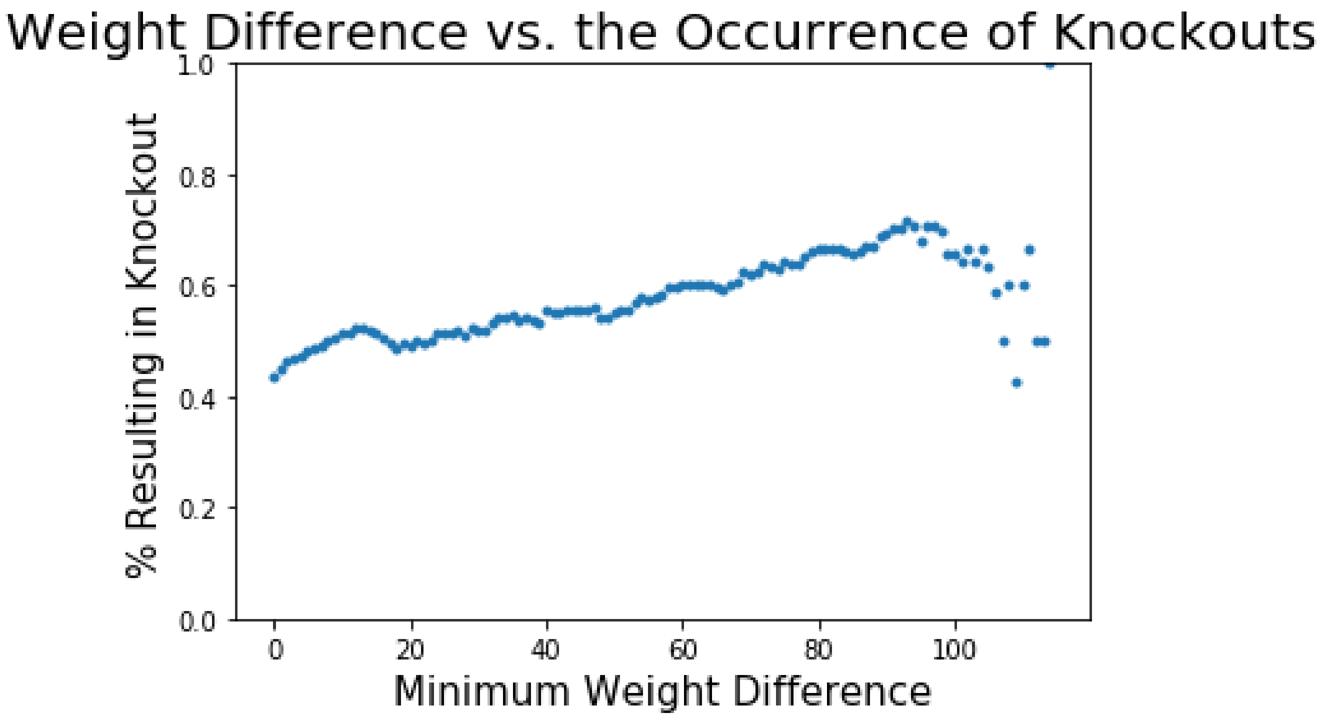
Predictive features so far

- Up until this point, we have seen that the differences in age, height, and reach are fairly good tools in predicting who won the match.
  - For matches where Boxer A won:
    - Mean diff_age = -2.1 years
    - Mean diff_height = 0.9 cm
    - Mean diff_weight = -0.2 lbs
    - Mean diff_reach = 1.1 cm

Another feature of interest

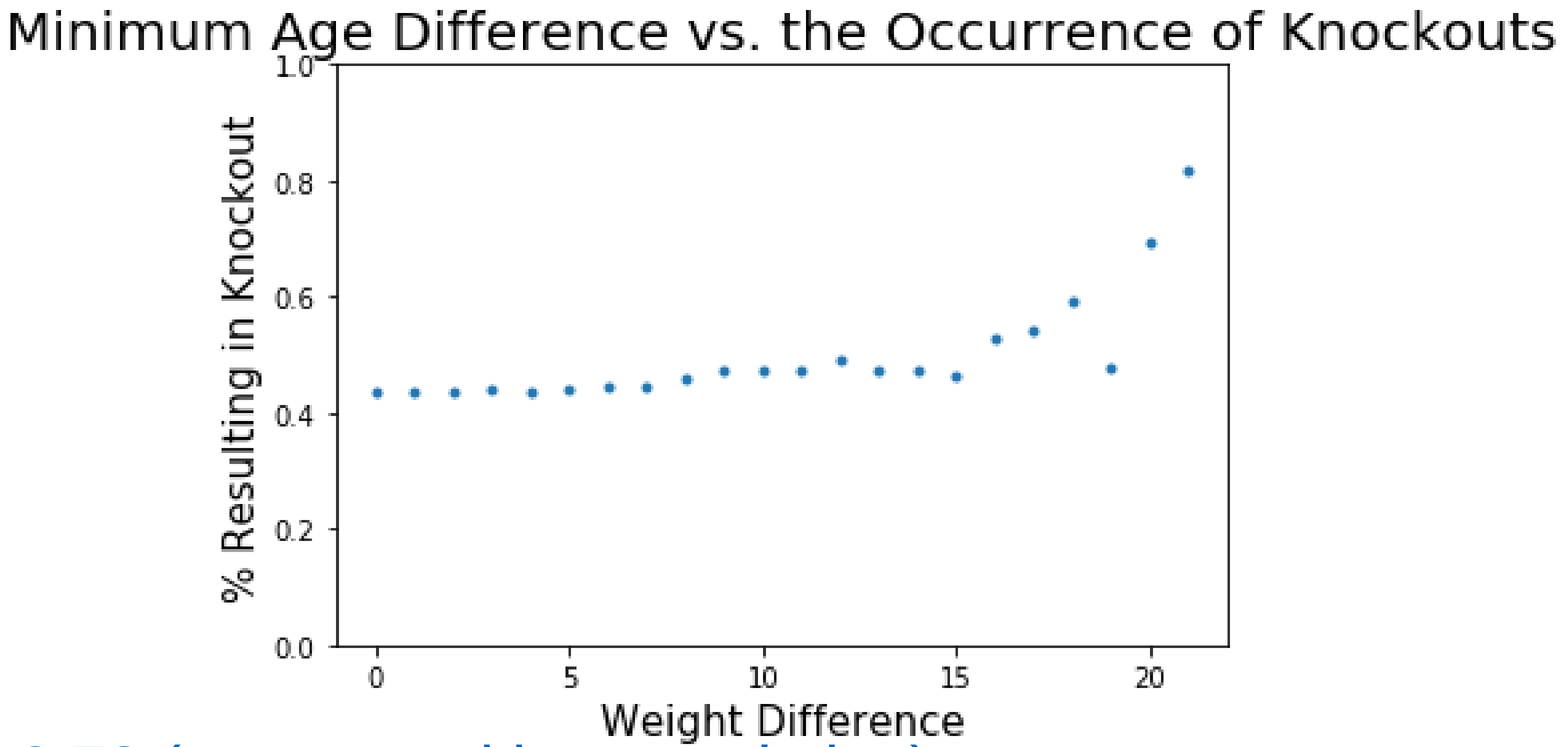
- We are interested in a popular metric of discussion in the boxing community: knockout percentage.
- So far we know that the differences in the physical attributes of the boxers has some relationship with the winner of the match.
  - Winning boxers tend to be younger, taller, lighter, and longer.
- What effect does the *size* of this disparity have on the result of the match?

diff_weight effect on K.O. likelihood



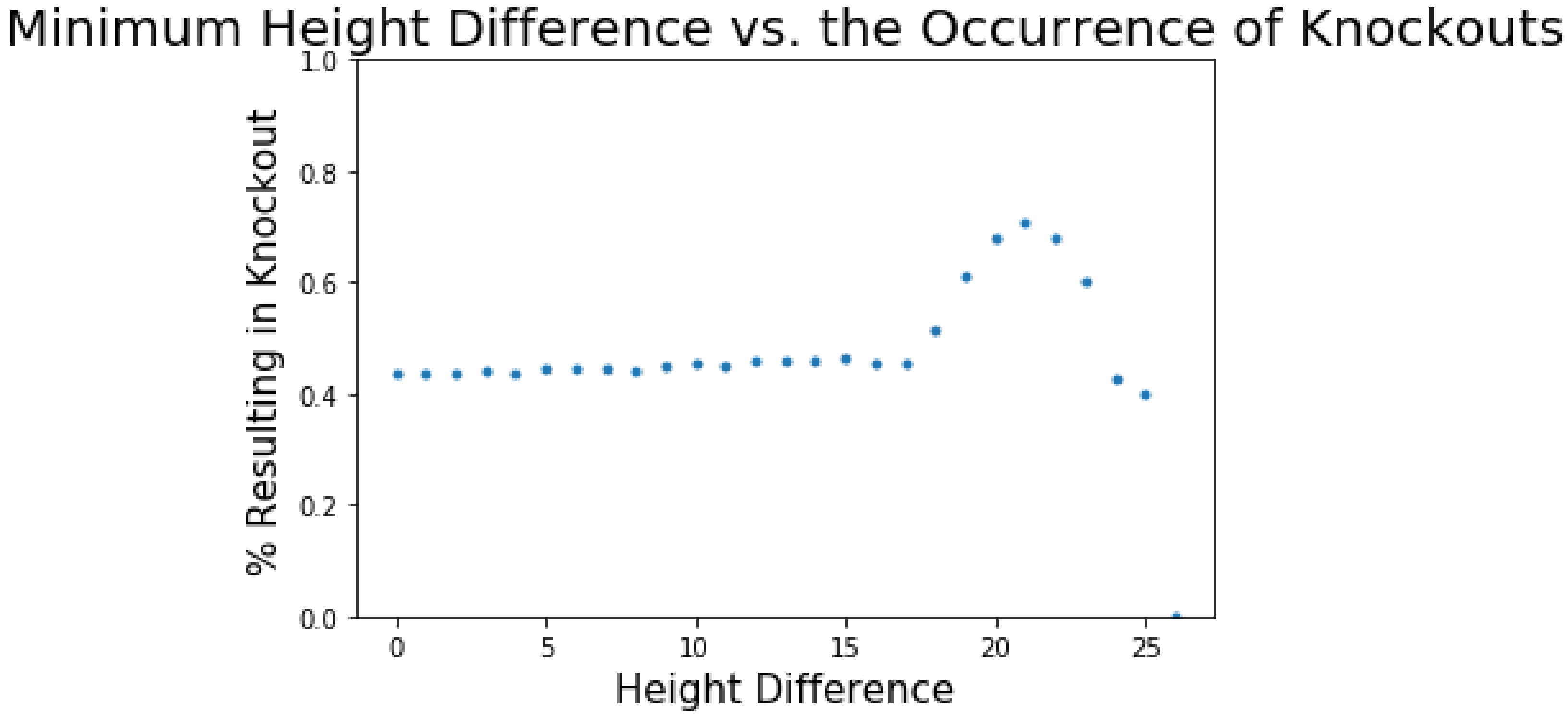
- r = 0.75 (strong positive correlation)
- As the disparity between the weights of the two boxers becomes larger, there is a clear increase in the percentage of matches resulting in knockout.
- We suspect that it might be easier for a heavier boxer to knock out a significantly lighter opponent.

diff_age effect on K.O. likelihood



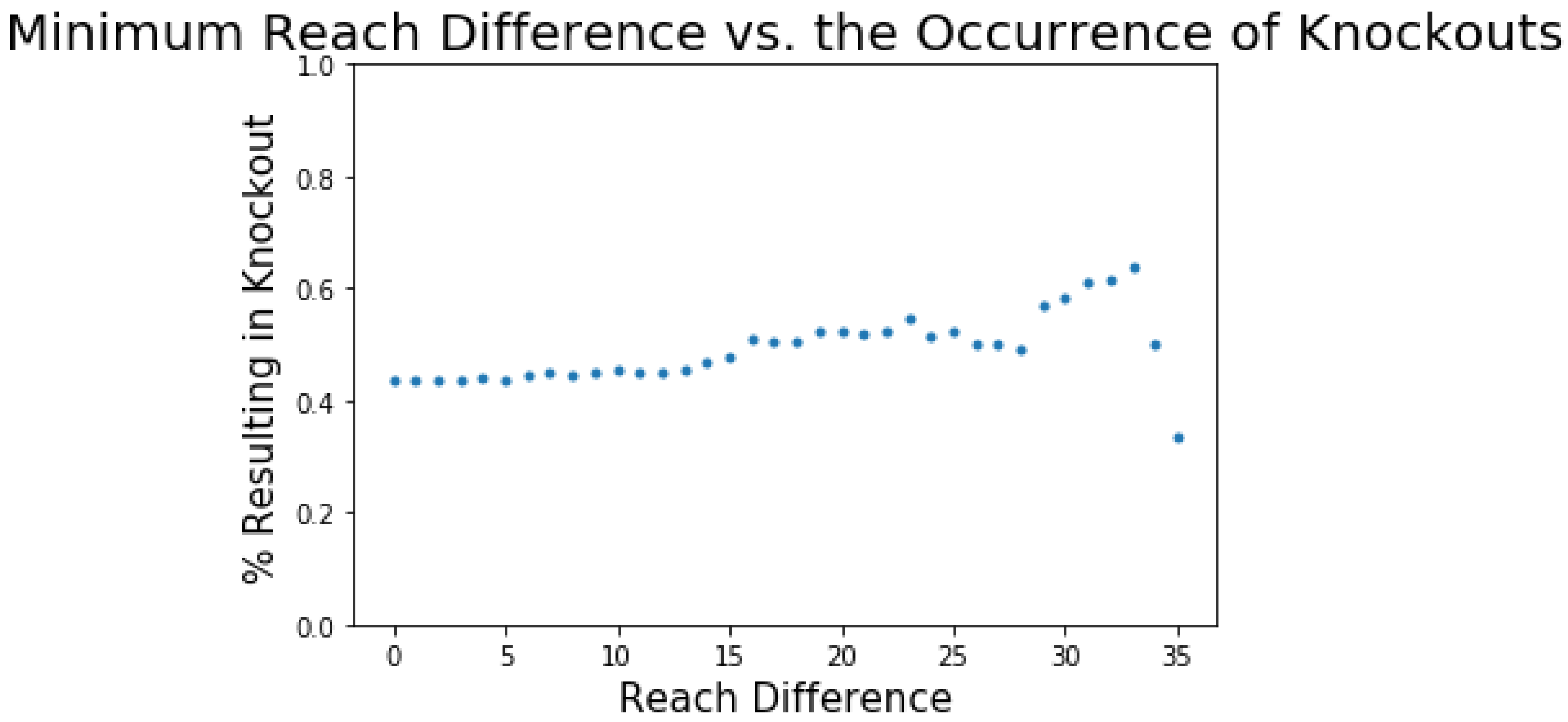
- r = 0.73 (strong positive correlation)
- Here the most significant uptake in knockout occurrence occurs when the age difference reaches approximately 15 years.
- The trend is once again upward.

diff_height effect on K.O. likelihood



- r = 0.10 (very weak correlation)
- Here there is no apparent trend. Overall, the size of the disparity in height between boxers does not seem to effect the likelihood of a knockout.

diff_reach effect on K.O. likelihood



- r = 0.62 (reasonably strong positive correlation)
- With the exception of the last two data points, we see a steady increase in knockout likelihood as we increase the disparity between the reaches of boxers.

Are the correlations observed significant?

- Previously, we tested if the distributions for the differences in physical attributes were different based on the winner of the match.
- Now we will test if the correlations observed between the size of these differences and the occurrence of knockouts are statistically significant.

Are the correlations significant?

- Each differential column tested using a correlation hypothesis test:
  - H_0: There is no correlation between the size of the difference and the percentage of matches resulting in knockout.
  - H_A: There is in fact a correlation.
- Again, very small p-values allow us to conclude that our observed correlations were significant.

Are the correlations significant?

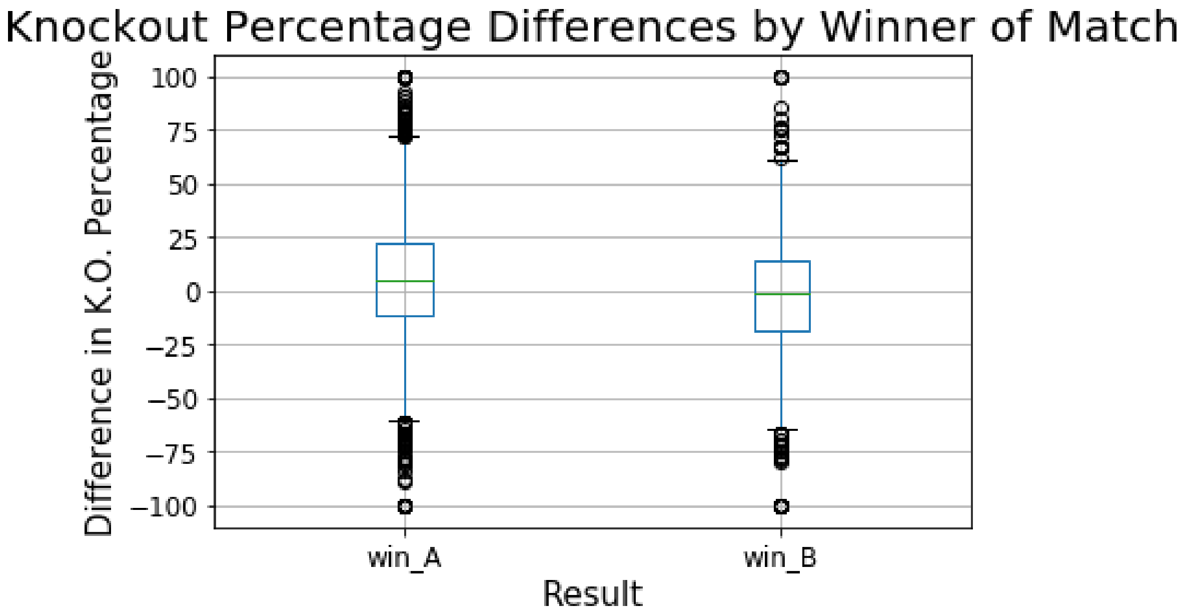
- Summary of hypothesis tests:
  - diff_weight → statistically significant
  - diff_age → statistically significant
  - diff_height → NOT statistically significant
    - This is not surprising, since our observed correlation coefficient was a very weak 0.10.
  - diff_reach → statistically significant

New predictive feature

- Because of these results, we are now interested in a new predictive feature: the difference in knockout percentages.
- We might expect that since more knockouts occur when the disparity in physical attributes is large, that a winning boxer might tend to have a higher knockout percentage than his opponent.

diff_ko_percentage EDA

- Our expectations are confirmed. More than 50% of the time, when Boxer A wins, he has a higher knockout percentage than his opponent.
- For matches where Boxer A won:
  - Mean diff_ko_percentage = 5.9%



Hypothesis test on diff_ko_percentage

- As before, we wish to test if the differences in knockout percentage when Boxer A wins are statistically different than when Boxer B wins.
- Using the same permutation hypothesis testing we used for the other four columns, we found that the differences in the two distributions are in fact statistically significant.

Our final set of predictive features

- We now have four fairly strong indicators of the winner of each match:
  - diff_age, diff_height, diff_reach, diff_ko_percentage
  - Again, we are excluding diff_weight, because there were not statistically significant differences between when Boxer A won and when Boxer B won.

Predicting the winner

- It is time to create a model to predict the winner of the match.
- Three candidate models:
  - Logistic Regression
  - Decision Tree
  - Support Vector Classifier

Steps in building a model

- Since there were far more winning Boxer A's than Boxer B's, we will use over-sampling to even out the imbalance.
- Split the data into two parts: a training set used to train the model, and a testing set that the model has never seen.
- Analyze the accuracy, precision, recall, and F1 scores of each model.

Best Logistic Regression model

- Accuracy: 63.7%

	Precision	Recall	F1 score	Support
Boxer A	0.63	0.64	0.64	1744
Boxer B	0.64	0.63	0.64	1775

Best Decision Tree model

- Accuracy: 91.9%

	Precision	Recall	F1 score	Support
Boxer A	0.99	0.85	0.91	1744
Boxer B	0.87	0.99	0.93	1775

Best model overall: Support Vector Classifier

- Accuracy: 99.3%

	Precision	Recall	F1 score	Support
Boxer A	0.99	1.00	0.99	1744
Boxer B	1.00	0.99	0.99	1775

Summary of model results

- Our best model overall was the support vector classifier, using RandomOverSampling to even out the class imbalance.
- Our four predictive features, diff_age, diff_height, diff_reach, and diff_ko_percentage, were highly effectively in predicting the winner of the match.

Further research

- What else can we explore?
  - Can we alter our model to predict a different target, such as the continuous values contained in the judges' scoring columns?
  - Could we predict a target variable with more than two classes, such as how the match ended (knockout, judges' decision, etc.)?
  - Could we include more or fewer features to improve our model?
  - What other information might we be able to gather about the boxers that would strengthen our predictive capabilities? Nationality? Diet? Number of years experience as a professional boxer?