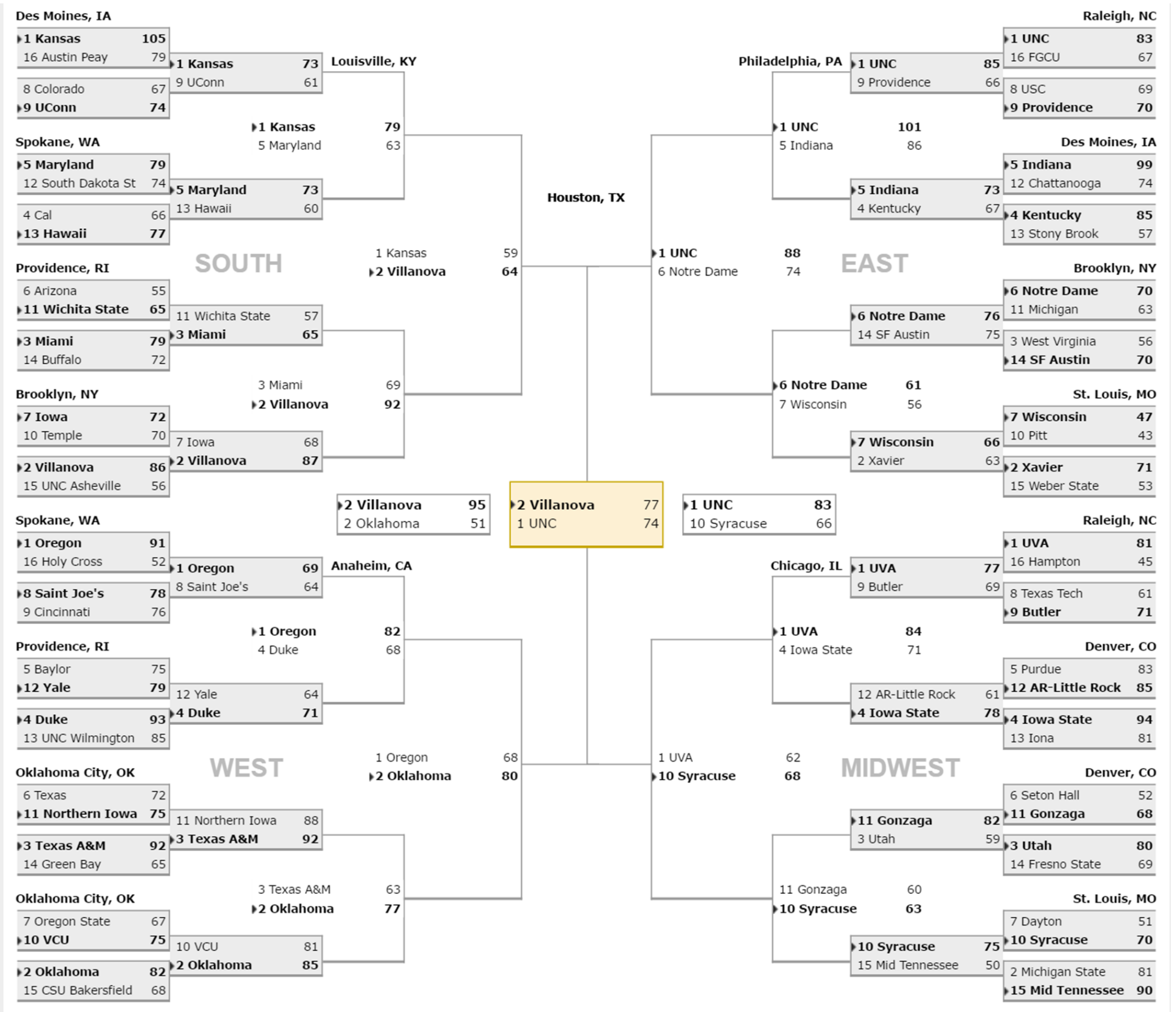
Predictive Model of 2016 March Madness Tournament

Authors: Bryan Katterman and Alejandro Martinez

Background

Relating the outcomes of sports with games of chance remains a complicated area of study in probability. By random chance, the odds of correctly predicting the outcome of all 63 games of the NCAA Basketball Tournament is 1 in 9,223,372,036,854,775,808 (over nine quintillion), however, the outcome in sports games almost never have fair odds; algorithms and weighting factors can be applied to help to narrow odds. The complexity in these weighted models is the enormous number of factors and data collection involved in accurately predicting a winner. Huge monetary prizes have been offered to anyone who can correctly guess the outcome of the tournament, such as Warren Buffet's 2014 $1-billion-dollar bracket prize as well as prizes offered by Quicken Loans and others. Our challenge was to gather statistics and data, create algorithms and weighting factors to create our own stochastic model .



http://espn.go.com/mens-college-basketball/tournament/bracket

Methods

In constructing the tournament prediction model, we gathered season win-percentage rates (GW), percentage of away games won (AW), and win rates for the 10 games prior to the tournament (Rf) on all 64 of the 2016 tournament teams and assigned weighting factors (W) to each of these statistics to create a weighted score for each of the participating teams.

$$Ws = W_1(GW\%) + W_2(AW\%) + W_3(Rf\%) \qquad \text{where} \qquad W_1 + W_2 + W_3 = 1$$

We also collected statistics on seed matchup performance (SMP) of each tournament from 1985 until 2015 (the entire history of the 64 game setup) to create additional weightings of the tournament outcome. With these basic statistics we researched statistical methods used by the sports industry in predicting odds and outcomes and decided on a method of using Bayesian inference, specifically using Bill James' 'Log5' algorithm derived from Bayes theorem as a base for our algorithm.

$$W\%_{A|B} = \frac{W\%_A(1 - W\%_B)}{W\%_A(1 - W\%_B) + W\%_b(1 - W\%_A)}$$

Using this base, we applied our weighted score for each team along with the seed matchup weightings

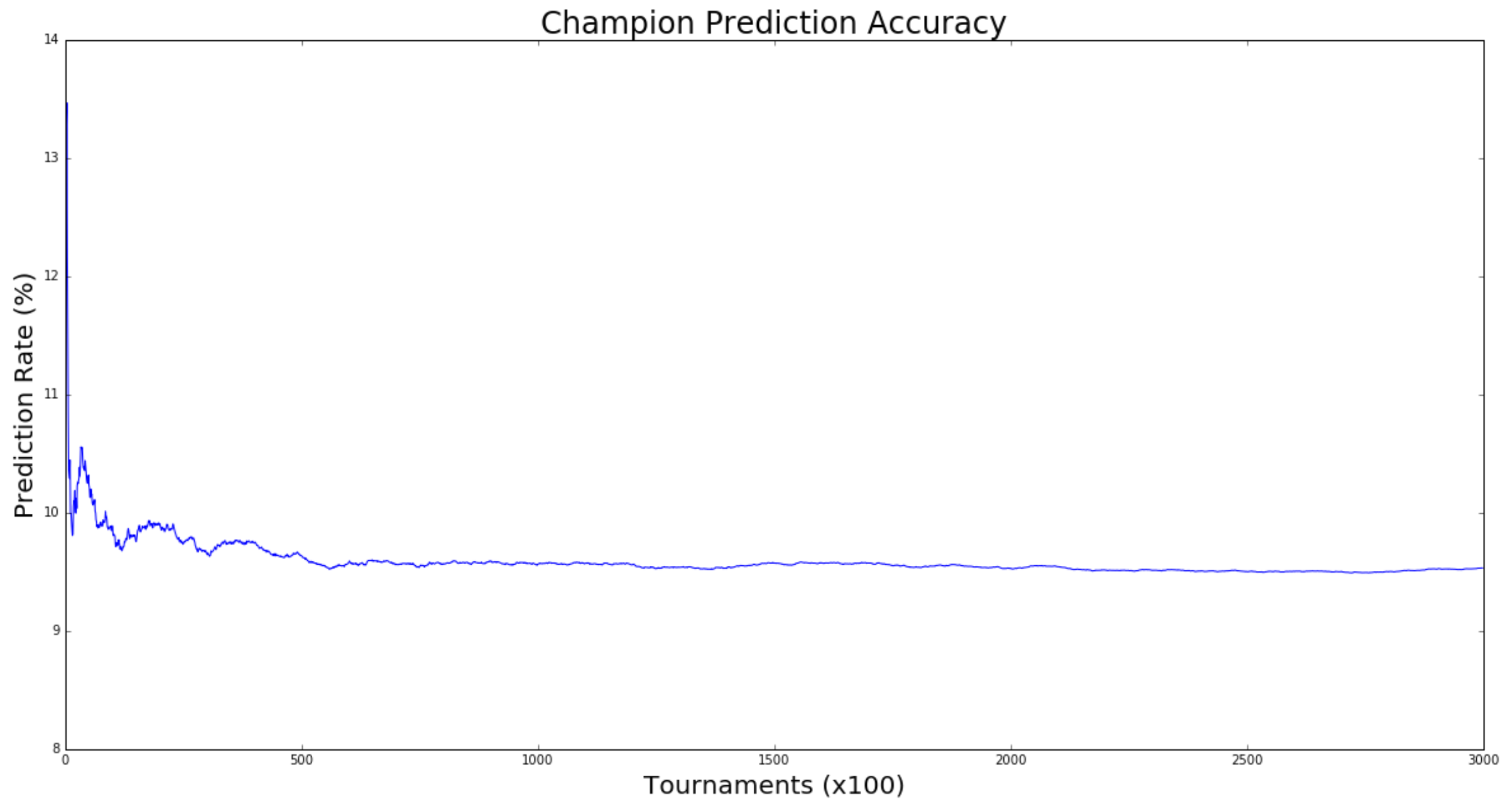
$$P(A_B) = \frac{Ws_A(1 - Ws_B)(SMP_A)}{Ws_A(1 - Ws_B)(SMP_A) + Ws_b(1 - Ws_A)(SMP_B)}$$

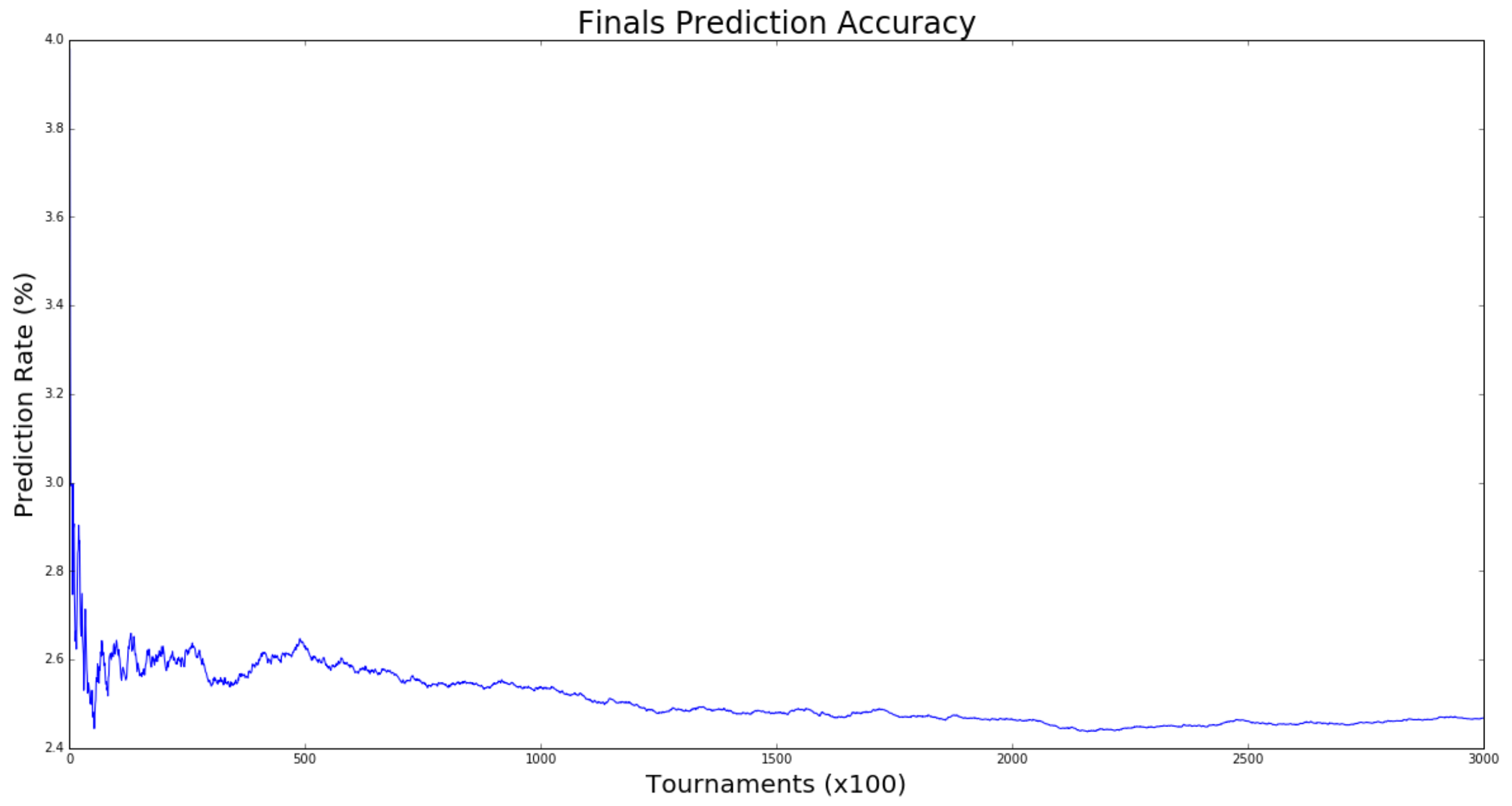
Even though there was 31 years of data to refer to, seed matchups data becomes more randomized the further you get into the tournament and therefore the data sets regarding seed matchup outcomes becomes statistically less reliable. So after the first round we dropped the seed matchup weighting.

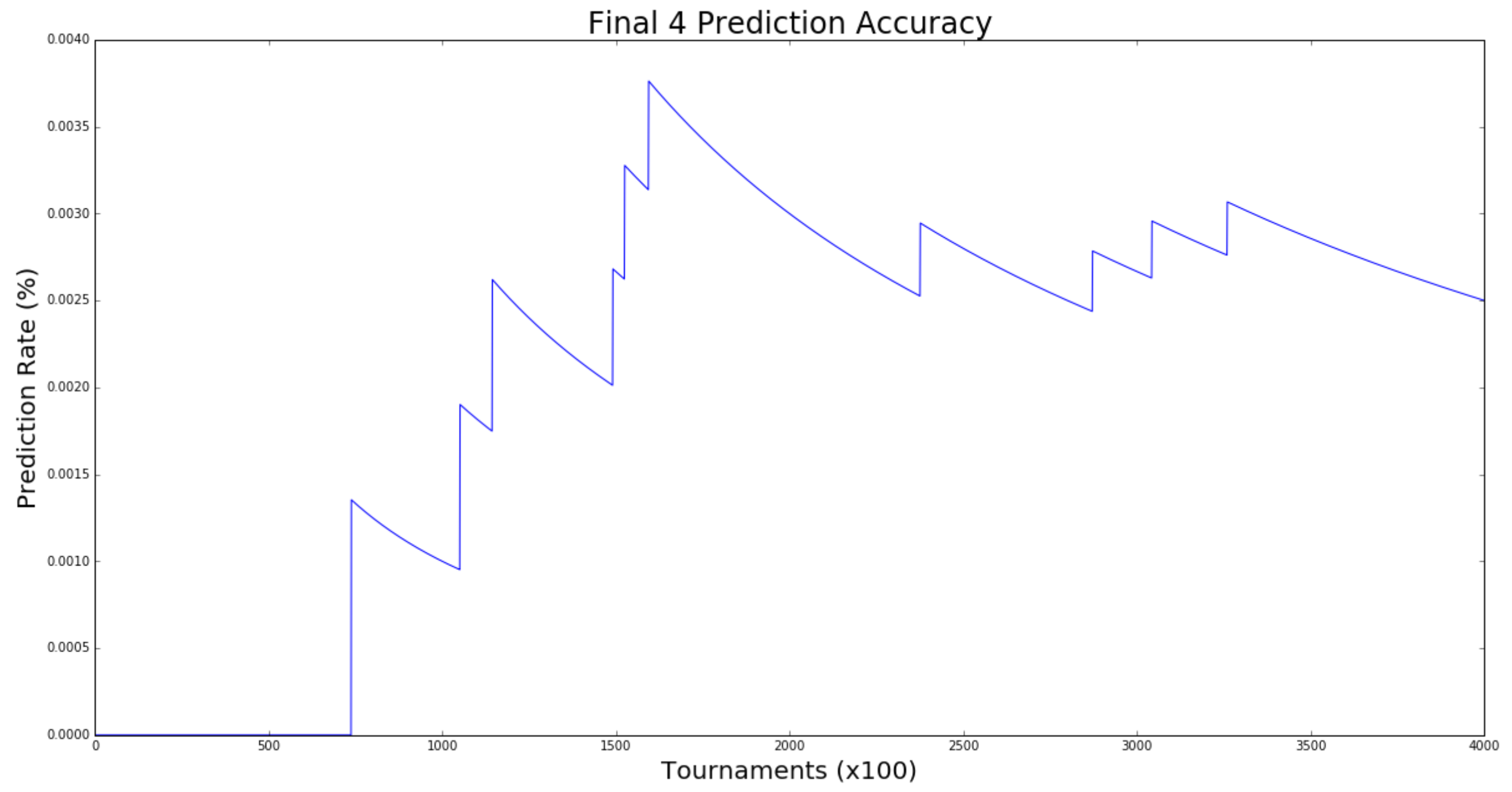
$$P(A_B) = \frac{Ws_A(1 - Ws_B)}{Ws_A(1 - Ws_B) + Ws_b(1 - Ws_A)}$$

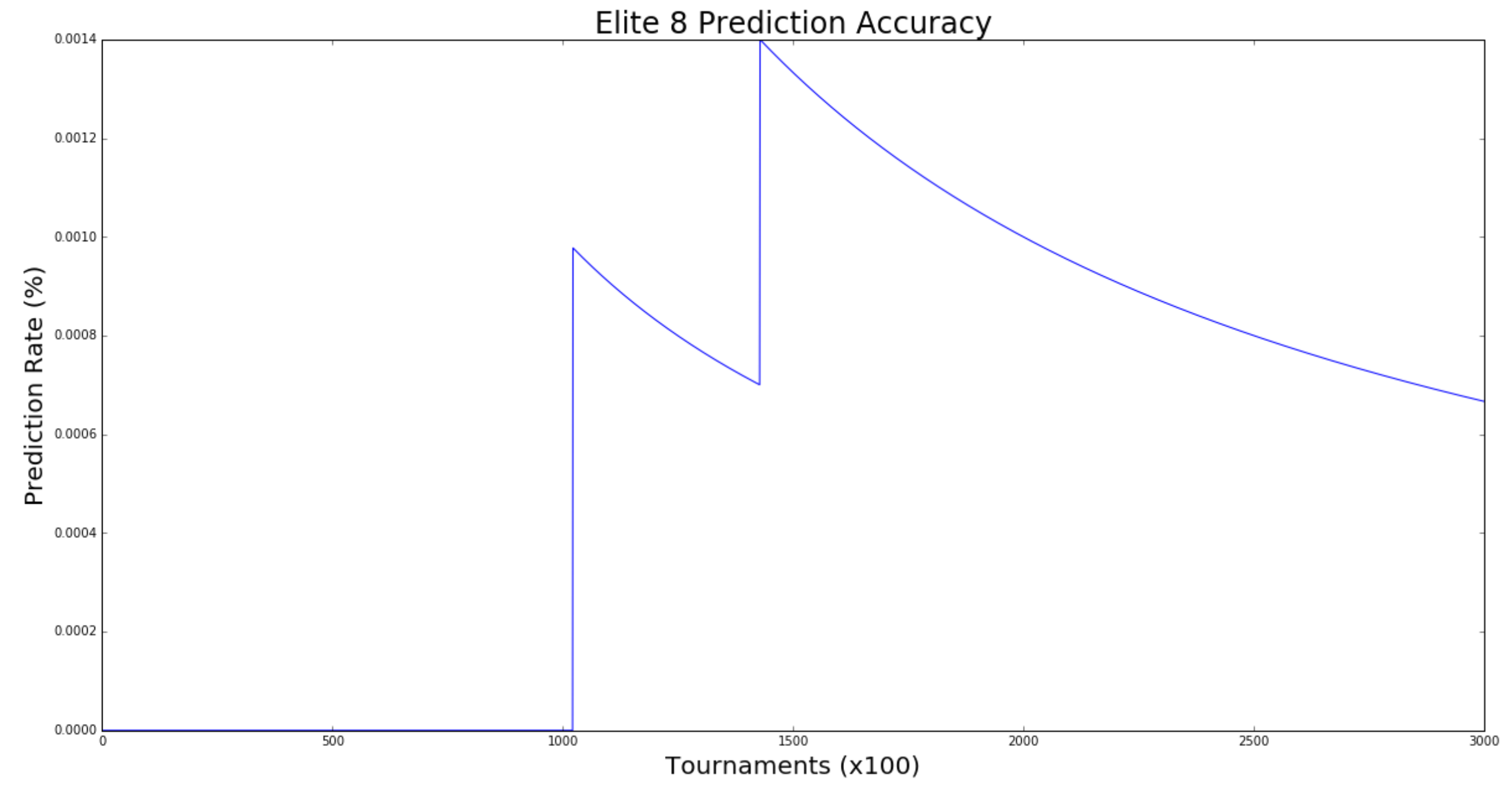
From there, create a Monte Carlo model using a random number generator to produce a number between 0 and 1. Using our probability algorithm, since $P(A_B) + P(B_A) = 1$ the winner of each round was selected if random number < P(A), then team A won. Otherwise team B moved on. Since enormous odd are involved in predicting all 63 games of the tournament correctly, we ran the Monte Carlo model to simulated hundreds of thousands, to 100 million of tournaments per set and analyzed the data.

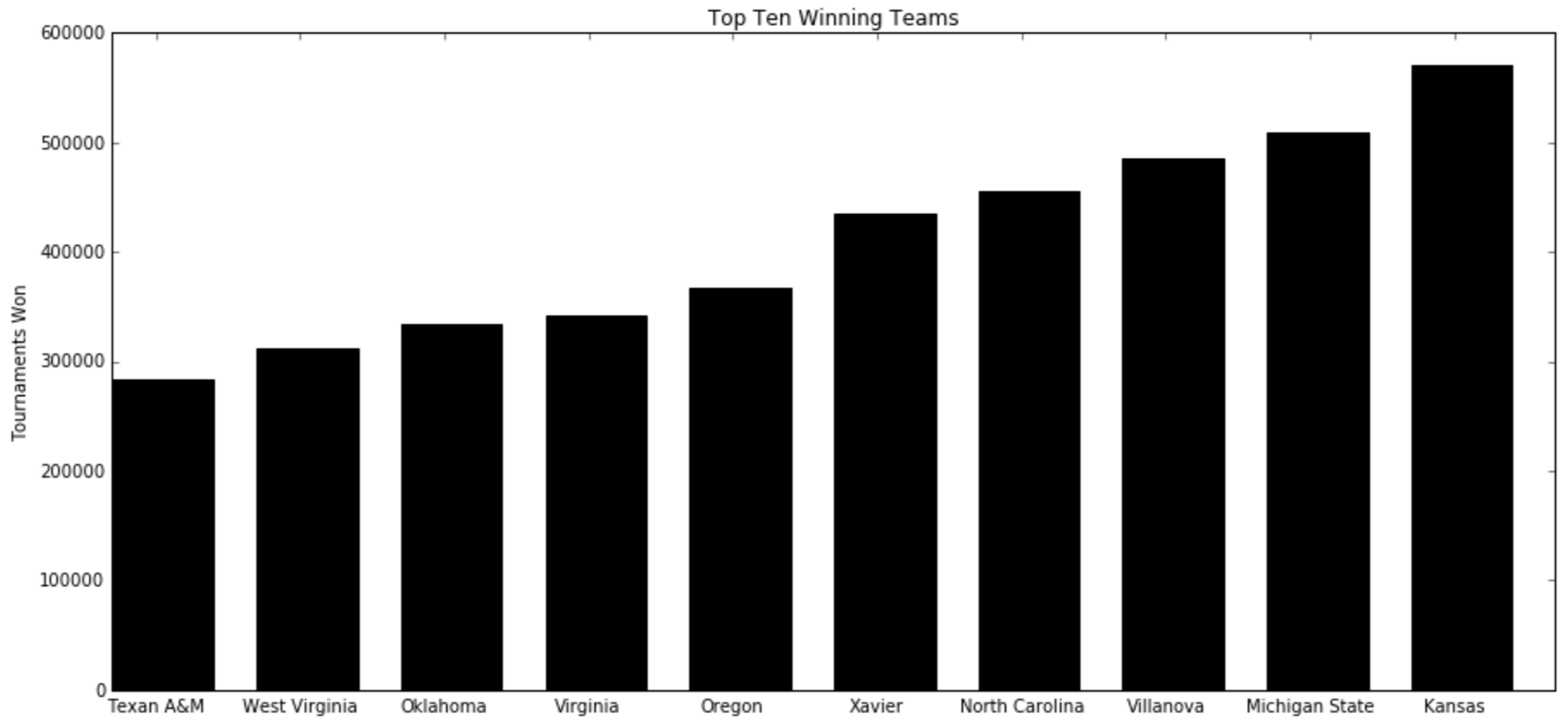
Results

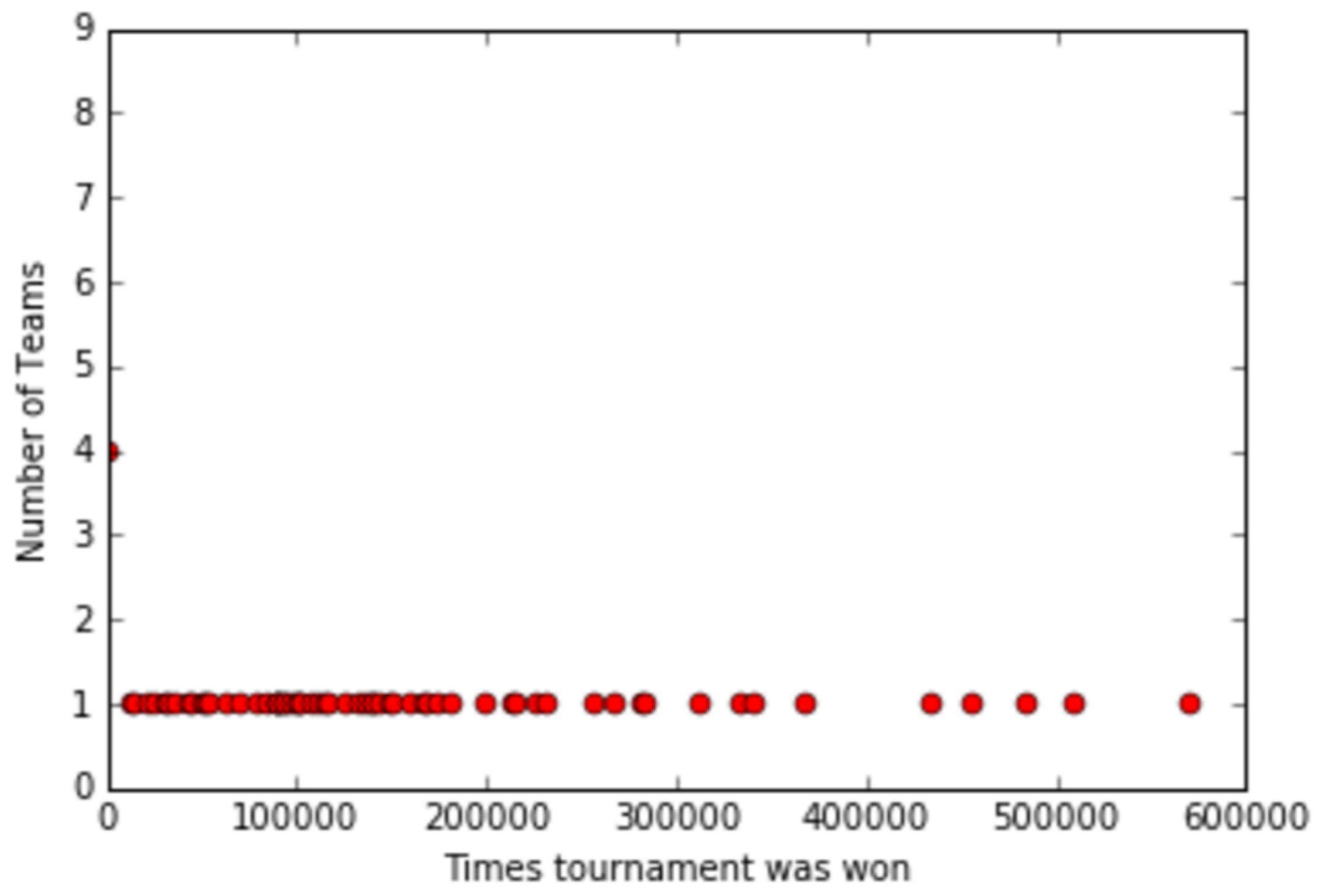












Summary

Overall, we achieved our goal of creating a predictive model using team statistics with varying results. Our model does fairly well with high ranking teams, but lacks uncertainty weightings to adequately predict low ranking teams that make it far in the tournament.

We plan on continuing our work on this model in an amateur capacity. Our plans for future work include optimizing weighting factors, gathering and implementing more data sets into the model including coaching and player stats, as well as analyzing our model with previous years tournaments. Our model was able to correctly predict the winner of the tournament at a rate of about 1 in 10.5 (about 9.6% of the time); the 2 teams that made it to the championship game at a rate of about 1 in 40.6 (about 2.5% of the time); the four teams that made it to the Final 4 about 1 in 45,000 (about 0.002% of the time); the eight teams that made it to the Elite 8 about 1 in 167,000 (about 0.0006% of the time) which are odds that are about 6, 26, 1.5 and 100 times, respectively , the odds of guessing the winners by random chance. The difference in the gap in accuracy between the Final 4 and final games is due to a low ranking team making it to the Final 4 this year (2nd lowest ranking team in history to make it) . Our model heavily favors high ranking teams and this reflects strongly in the comparative analysis.

References

Miller, Steven J. "A JUSTIFICATION OF THE log 5 RULE FOR WINNING PERCENTAGES." (2008).

espn.com

apbr.org

Code is available at https://github.com/ASU-CompMethodsPhysics-PHY494/final-march-madness.git [under<Apache 2.0>]