Wharton University of Pennsylvania SPORTS ANALYTICS Student Research Journal

2022, Vol 4

wsb.wharton.upenn.edu/student-research-journal

# Predicting Winners in Cricket: introducing sports analytics to one of the world's most watched sports

***Amy Liu, W'23***
***The Wharton School of the University of Pennsylvania, Philadelphia PA, USA***

***Sardar A. Cheema, W'21***
***The Wharton School of the University of Pennsylvania, Philadelphia PA, USA***

**Advisor: Abraham Wyner, PhD**
***The Wharton School of the University of Pennsylvania, Philadelphia PA, USA***

**Abstract**

The rise in sports analytics has focused on popular American games such as baseball and basketball, as well as soccer. While cricket is one of the most popular sports on a global level, there has been limited application of sport analytics reported in the literature. In this work, established sports analytics methods are applied to cricket; two cricket models are created, a Pythagorean win rate, and an Elo rating system. The cricket Pythagorean win rate was built from One Day Internationals (ODI) results; the runs scored and the runs conceded on a per-year basis for each team, for data spanning the years 2000 to 2021. A limitation of a Pythagorean win rate is that it does not account for the quality of the competing team, and hence an ELO model was also created. The ELO model was built from International Cricket Council (ICC) full members results in ODIs. The Elo model was initialized with ratings = 1500 at the start of the first year, the model then calculated the ratings for all 27 teams in the dataset. Plotting the change over time made it obvious that we should only look at the full members and remove most of the associate members since there are very few data points for the latter. While to date cricket has not followed the movement towards data analytics seen in other major sports, our work shows there are no inherent limitations in cricket data. Both models created are decently robust, and we hope that they will serve as a foundation for others to build on. We hope this works starts a conversation about analytics in cricket.

Keywords: cricket, ELO, Pythagorean win, One Day Internationals, win prediction

## Introduction

Sports analytics, a burgeoning area of study, has shown us that data-oriented thinking can be applied to any conceivable purpose. Sports analytics studies and experiments have traditionallyfocused on popular all-American games such as baseball and basketball. Our goal with the finalproject was to apply established sports analytics methods to cricket, where little exploration hasbeen done in the existing literature. In this report we are creating two models, one for the Pythagorean win rate in the context of cricket, and a second to develop an Elo rating system.

Before we delve deeper into the data collection and model building process, it is important to explain a few concepts about the sport itself. There are three main formats in cricket: (1) Test cricket, (2) One Day Internationals and, (3) Twenty20. Test cricket is the oldest and longest format where the game does not end in a day and is usually decided across 4 or 5 days and more than one innings is played by each team. The latter two are collectively called 'limited overs' cricket and limited over games end in a single day. In this paper, we are looking at One Day International cricket (ODI, from this point onwards) to develop our models.

The format of the game is simple. There are 11 players on each team, this includes batsmen, bowlers and all-rounders (players who can bat and bowl). The usual combination a team goes with is 4 bowlers and 7 batsmen/all-rounders. It is important to achieve a good balance between overall batting and bowling strength as the same 11 players are expected to play both innings. The game starts with a toss and whoever wins the toss decides to bat or fieldfirst. When Team A is batting, Team B fields/bowls. At the end of the innings, they switch roles and the second innings decides who wins. Let's suppose that Team A wins and decides to bat first. At any given point, 2 players from Team A and all 11 players from Team B are on the field.The objective for Team A is to get as many runs possible in 50 overs (6 balls are bowled per over). The objective for Team B is to restrict Team A to as low a score as possible. For brevity'ssake, we will not go in depth over how runs are scored and how players get out but it is not very different from baseball (think baseball with 2 bases instead of 4).

The last bit of information that is important to know about cricket is the concept of ICC full members and associate members. The ICC is the International Cricket Council and is the overarching governing body for international cricket. Full members are countries recognized byICC as official test match playing countries and they each have full voting rights at ICC meetings. In simpler terms, they are the older, more established cricket playing nations. Associate members (also in simpler terms) are new/emerging members that are either or not asexperienced or do not devote as many resources towards the sport to build a strong international team.

## Data collection

Since there are no repositories where one can access compiled cricket data, we had to scrapewhat we needed form the following website:
https://stats.espncricinfo.com/ci/content/records/307851.html

Here, each link represents the ODI matches played for a given year (used interchangeably withseason in this paper). Our algorithm went inside each link to arrive at the following:

| Team 1 | Team 2 | Winner | Margin | Ground | Match Date | Scorecard |
|---|---|---|---|---|---|---|
| New Zealand | West Indies | New Zealand | 3 wickets | Auckland | Jan 2, 2000 | ODI # 1532 |
| New Zealand | West Indies | New Zealand | 7 wickets | Taupo | Jan 4, 2000 | ODI # 1533 |
| New Zealand | West Indies | New Zealand | 4 wickets | Napier | Jan 6, 2000 | ODI # 1534 |
| New Zealand | West Indies | New Zealand | 8 wickets | Wellington | Jan 8-9, 2000 | ODI # 1535 |
| Australia | Pakistan | Pakistan | 45 runs | Brisbane | Jan 9, 2000 | ODI # 1536 |
| India | Pakistan | Pakistan | 2 wickets | Brisbane | Jan 10, 2000 | ODI # 1537 |
| New Zealand | West Indies | New Zealand | 20 runs | Christchurch | Jan 11, 2000 | ODI # 1538 |
| Australia | India | Australia | 28 runs | Melbourne | Jan 12, 2000 | ODI # 1539 |
| Australia | India | Australia | 5 wickets | Sydney | Jan 14, 2000 | ODI # 1540 |
| Australia | Pakistan | Australia | 6 wickets | Melbourne | Jan 16, 2000 | ODI # 1541 |
| Australia | Pakistan | Australia | 81 runs | Sydney | Jan 19, 2000 | ODI # 1542 |
| India | Pakistan | Pakistan | 32 runs | Hobart | Jan 21, 2000 | ODI # 1543 |
| South Africa | Zimbabwe | South Africa | 6 wickets | Johannesburg | Jan 21, 2000 | ODI # 1544 |
| Australia | Pakistan | Australia | 15 runs | Melbourne | Jan 23, 2000 | ODI # 1545 |
| South Africa | England | England | 9 wickets | Bloemfontein | Jan 23, 2000 | ODI # 1546 |
| India | Pakistan | India | 48 runs | Adelaide | Jan 25, 2000 | ODI # 1547 |
| Australia | India | Australia | 152 runs | Adelaide | Jan 26, 2000 | ODI # 1548 |
| South Africa | England | South Africa | 1 run | Cape Town | Jan 26, 2000 | ODI # 1549 |
| India | Pakistan | Pakistan | 104 runs | Perth | Jan 28, 2000 | ODI # 1550 |
| England | Zimbabwe | Zimbabwe | 104 runs | Cape Town | Jan 28, 2000 | ODI # 1551 |
| Australia | India | Australia | 4 wickets | Perth | Jan 30, 2000 | ODI # 1552 |
| England | Zimbabwe | England | 8 wickets | Kimberley | Jan 30, 2000 | ODI # 1553 |
| Australia | Pakistan | Australia | 6 wickets | Melbourne | Feb 2, 2000 | ODI # 1554 |
| South Africa | Zimbabwe | Zimbabwe | 2 wickets | Durban | Feb 2, 2000 | ODI # 1555 |
| Australia | Pakistan | Australia | 152 runs | Sydney | Feb 4, 2000 | ODI # 1556 |
| South Africa | England | South Africa | 2 wickets | East London | Feb 4, 2000 | ODI # 1557 |
| South Africa | Zimbabwe | South Africa | 53 runs | Port Elizabeth | Feb 6, 2000 | ODI # 1558 |
| Pakistan | Sri Lanka | Sri Lanka | 29 runs | Karachi | Feb 13, 2000 | ODI # 1559 |
| South Africa | England | South Africa | 38 runs | Johannesburg | Feb 13, 2000 | ODI # 1560 |
| Pakistan | Sri Lanka | Sri Lanka | 34 runs | Gujranwala | Feb 16, 2000 | ODI # 1561 |
| Zimbabwe | England | England | 5 wickets | Bulawayo | Feb 16, 2000 | ODI # 1562 |
| New Zealand | Australia | no result | | Wellington | Feb 17, 2000 | ODI # 1563 |
| Zimbabwe | England | England | 1 wicket | Bulawayo | Feb 18, 2000 | ODI # 1564 |
| New Zealand | Australia | Australia | 5 wickets | Auckland | Feb 19, 2000 | ODI # 1565 |
| Pakistan | Sri Lanka | Sri Lanka | 104 runs | Lahore | Feb 19, 2000 | ODI # 1566 |

This is the data that was subsequently scraped and serves as the basis of our analysis. To get the scores for the individual teams, we wrote a script to click on the individual ODI tags in the scorecard column to access the page for that specific match and scrape the scores for the teams. The data that we scraped is for every ODI match that was played between 2000 up until early 2021 and it includes 5000+ records for matches played by 27 teams. We chose not to explain the cleaning process in detail as it does not add much value to the paper.

**Pythagorean win rate**

Because not much analysis has historically been done with cricket data, any data we could collect started in its most unpolished and raw form. For example, inconsistent scorekeeping methods in the same columns, extremely granular and game-specific data rather than more high-level summarizing datasets, the random scatter of important information across different datasets, and other detailed issues required intensive cleaning, reshaping, and data merging.

Branching off from Project 1, our preliminary analysis was to determine a Pythagorean win rate for cricket. However, now we are doing so for ODI cricket instead of T-20 cricket. Win rate exponents have been established and agreed upon for other sports such as Major League Baseball (1.8 or 2), but there is currently no widely accepted exponent for the Pythagorean calculator for cricket. We use the same core formula, equation (1):

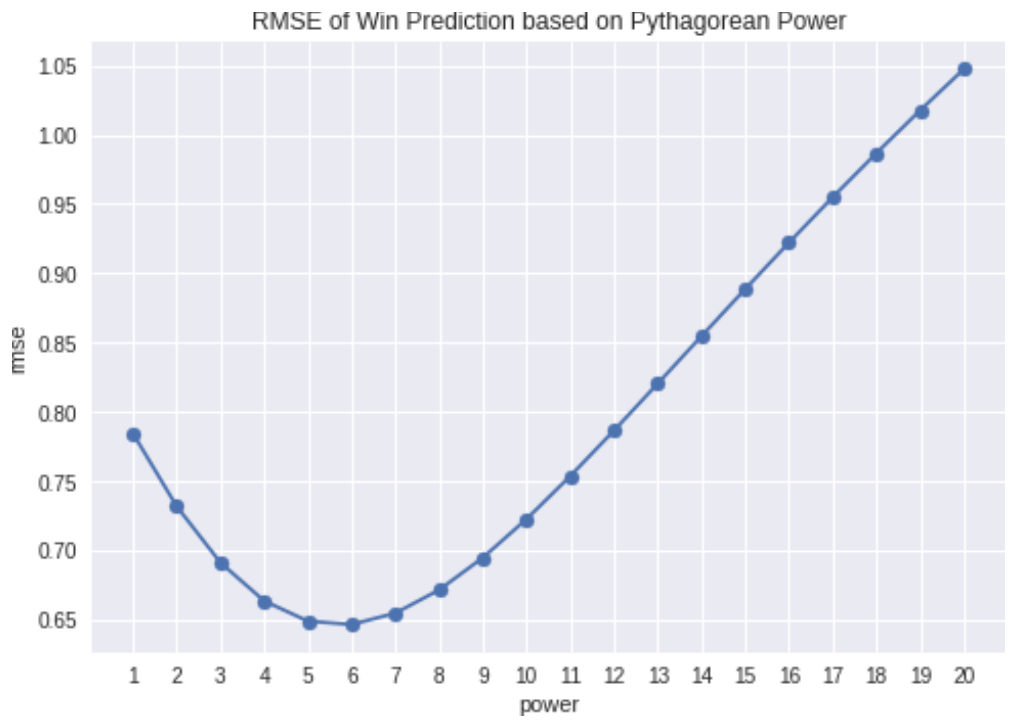
(1) $$WinPC = \frac{Points\ Scored^k}{Points\ Scored^k + Points\ Conceded^k}$$

Which is derived from regressing the natural log of the win odds over the natural log of the runs (points)-scored-to-runs-allowed ratio. Through some arithmetic manipulations, the correlation coefficient of this regression would become the exponent of this win rate formula, k. In our context, we accumulate the runs scored and the runs conceded on a per-year basis for each team, for data spanning the years 2000 to 2021. We randomly selected 70 percent of the rows as our training set, and all remaining rows as our test dataset. Our dataset had thousands of observations, so we were confident we had enough data for training purposes. The random split ensured that the two groups would not differ by changes to country environment, team skill, or any other hidden factors correlated with time.

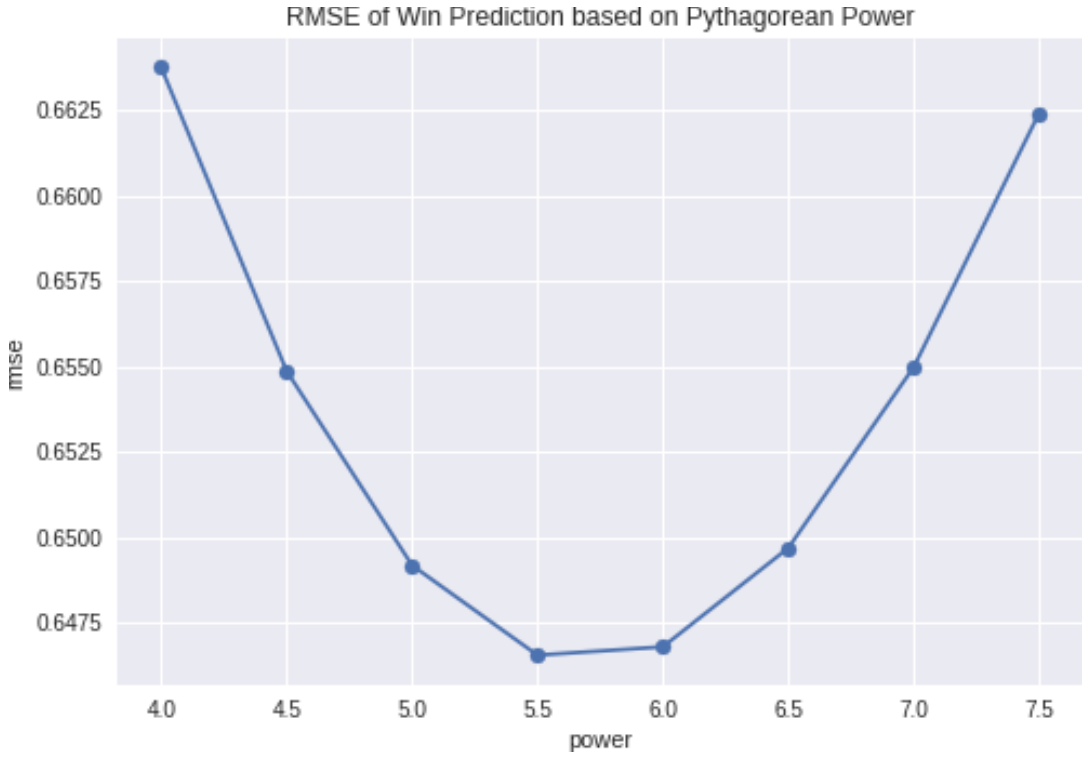
For each team and season, we calculated 20 different versions of the win PC ratio by inputting the total runs scored and conceded into the formula but varying the exponent k from 1 to 20. We then calculated the root mean square error for each of the 20 candidates by comparing the actual wins a given team earned that season to the modeled number of wins predicted by that value of k. The value of k that minimized the rmse was the exponent selected by our training data.

(2) $$Root\ Mean\ Square\ Error = \sqrt{\frac{\sum(Actual\ Wins - Predicted\ Wins)^2}{Toal\ Number\ of\ Matchers}}$$

For ODI cricket games, we determined an exponent of about 5.5. The minimized rmse valuewas 0.646. A potential issue we identified is that some teams in the dataset may have only played 2 or 3 games. Although we still wanted to include their information, we knew this could introduce variability into the predictions riding off only very few matches played. For example, Afghanistan won 3 out of 3 total matches played in 2021.




When using this exponent of **5.5** on our test dataset, our rmse value of **0.647** comes extremely close to the rmse of the training dataset of 0.646. However, all rmse values clustered between 4 and 7 end up being extremely similar (see graph on next page) and hovering around 0.64-0.65. The more important takeaway is that the test rmse values are similar to those of the training data and follow a similar trend. This means that the Pythagorean win rate method successfully avoids overfitting training data. Our model performs well out-of-sample and there is consistencybetween the two datasets.




Following this, we decided to calculate the optimal exponent value by maximizing the Log- likelihood between the actual winning percentage and the predicted winning percentage for each country per season. Doing so gave us an exponent of **4.94**. This exponent value, when used on the test dataset, gave us an rmse value of **0.642** which is slightly better than what wewere working with previously. Therefore, at the end of this model, we concluded that using anexponent value of 4.94 would give us the most accurate Pythagorean win rate for ODI cricket.

This value is slightly lower than the one we arrived at for T-20 cricket in Project 1 and this is a result of the difference in the two formats. T-20s are more fast paced games and see more runsscored per over, on average, than ODI cricket. Moreover, T-20 cricket matches last for about 3 hours while ODIs last for 8 hours. As a result, it is easier for batsmen to score more runs per ball as the game is much shorter. These factors contribute to the difference in score, winning totals and win margins which in turn lead to us getting different results for our Pythagorean model.

**Elo Model**

Seeing how the Pythagorean win rate does not account for the caliber and skill level of the teams that are competing against each other, it does not accurately reflect the odds of one teamwinning against the other in any future setting. Therefore, to create a model that does capture this idea of the teams having a difference in

skill level, we decided to build an Elo model. The base variables that form the foundation of the model are mean Elo rating, Elo width and K- factor. For the purposes of our project, it is not necessary to understand the technical details behind these three but here are some simple definitions as to what they represent:

1. Mean Elo rating: the mean around which the ratings deviate and the rating that the team start off with at the beginning of the time period
2. Elo width: how high or low the Elo ratings can go, i.e. what is the theoretical maximum and minimum rating a team can achieve
3. K-factor: the number that determines how quickly a rating reacts to the results of a newgame

Since we are building this model from scratch and do not have established cricket Elo models towork with, we decided to set the following:

1. Mean Elo rating = 1500
2. ELO width = 400
3. K-factor = 50

The mean Elo rating was fixed at 1500 because that appears to be the general mean for mostmodels and the same is true for the width. The k-factor, however, had to be adapted for the sport. While certain games like chess that see a lot of matches being played use sophisticatedmethods to assign k-factors such as the following:

3) $$K = \frac{800}{N_e + m}$$

where Ne is the number of games a player's rating is based on and m is the number of games aplayer completed in a tournament, we decided to simplify things and go down the football/baseball route where each game is assigned the same weight (or importance)[1]. The value we arrived at, k=50, allows us to give more importance to a game than other sports such as football (where k usually hovers between 20 and 40) because we do not see as many ODI series being played between teams when compared to the number of games played in a seasonfor football.

The model itself uses a fairly intuitive process to update the Elo ratings for two teams once theyhave played a match against each other. Firstly, the probability of one team winning against theother is calculated based on their current Elo ratings. The equation used here is:

4) $$Expected = \frac{1}{(1 + 10^{((loser_{elo} - winner_{elo})/elo\ width)}}$$

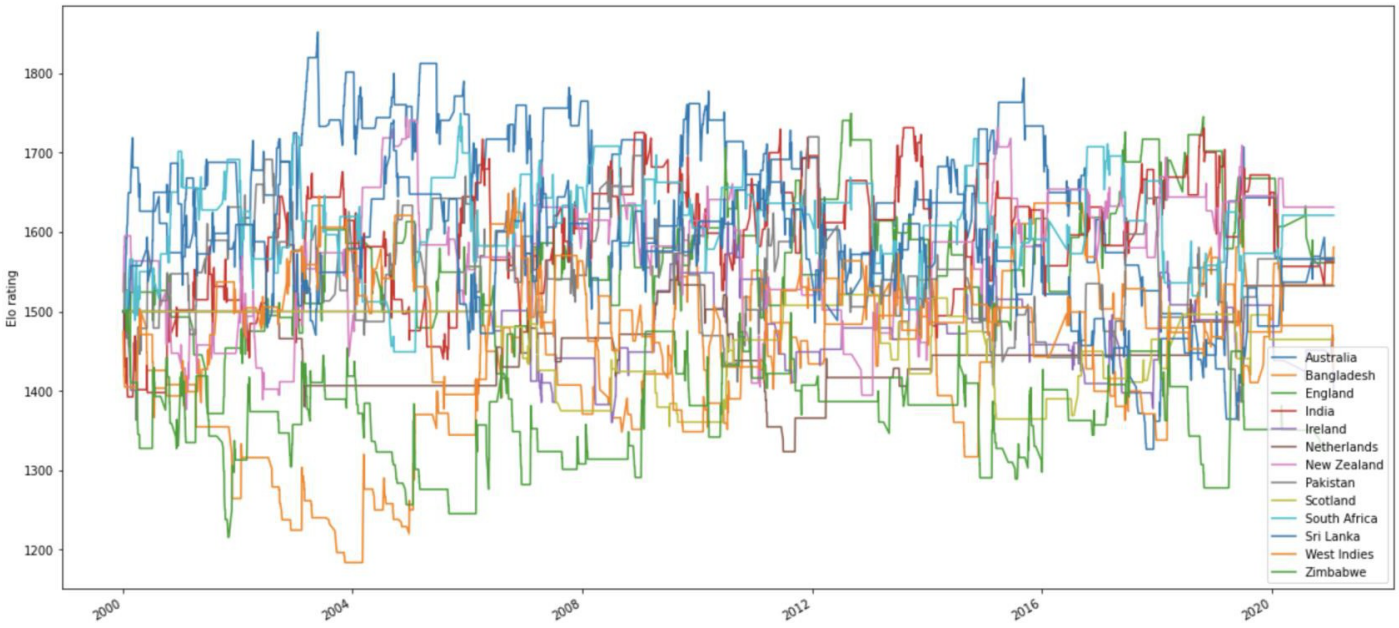
Based on the expected value, the Elo ratings for the teams are updated using the followingformula where the result is added to the rating of the winning team and subtracted from therating of the losing team:

(5) $$Change\ in\ Elo = K * (1 - Expected)$$

The last step is to regress the ratings towards the mean at the end of each year. This ensuresthat the ratings do not keep increasing towards infinity and vice versa. It also offers for a bettercomparison when comparing the different team ratings.

6) $$Elo_{t+1} = \frac{Elo_t - Elo_{mean}}{3}$$

After creating the Elo model and initializing it with ratings = 1500 at the start of the first year, we used the model to calculate the ratings for all 27 teams in the dataset. Plotting the change over time made it obvious that we should only look at the full members and remove most of the associate members since there are very few data points for the latter (refer to the last paragraphof the introduction for the difference between the two).




The final Elo ratings for the teams

**Limitations & Improvements**

The main limitation for the model is one that can be rectified and that is the time period beingconsidered. If we were to get data from the 1980s onwards, the ratings would be more representative. However, we would have to remove additional countries such as Bangladesh,Zimbabwe, Netherlands etc from the analysis as these teams are much younger when compared with the rest. The other thing that can be improved upon is the k-factor value.

Assigning different values based on the stage the match is being played at would offer a greaterdegree of sophistication as more important games such as World Cup matches will make a bigger impact on a team's

ratings than less important bilateral series between two countries.
Lastly, the model does not take into account the fact that batting and bowling conditions varydrastically depending on the pitch, weather, homeground advantage etc. so another improvement would be to include these factors when calculating the change in ratings.

## Conclusion

As we have highlighted time and again, sports analytics has not made strides in cricket the way it has in other sports such as football, baseball, basketball etc. This is not a product of it being more difficult to incorporate analytics into cricket. Our work has shown that it is very easy to do so. Unfortunately, cricket has not adapted a similar shift towards data analytics as other sports.We believe that the greatest purpose our paper can serve is to start a conversation about analytics in cricket. Both models that we created are decently robust with some limitations (as highlighted in each section) and we hope that they will serve as a foundation for others to build on.

## Reference

) https://staturdays.com/2020/08/11/introducing-college-football-elo-ratings/