Project Report on
Predicting Team Standings in ODI for a Cricket League

Anshul Roonwal
SBU ID: 110554783
aroonwal@cs.stonybrook.edu

Sahil Jain
SBU ID: 110281300
sahjain@cs.stonybrook.edu

1. INTRODUCTION

Originated in England, Cricket is a sport that plays an important part in day to day life of many people in Britain, South Africa, Australia, Indian Subcontinent and West Indies. This is the reason why it is the national sport of Australia, England and South Africa. It is one kind of outdoor game which is played between two teams with 11 players a side. It is played with a ball, a bat and wickets on an oval shaped ground.

There are different types of game formats. Typically, they fall in only three categories, ODI[1] (50, six balls over match), Test Cricket (a 5-day long match) and a T20 (with overs reduced to 20 instead of 50 in ODI[1]). There are popular leagues that takes place in all of these three formats. Some popular ones are – Ashes (played between England and Australia), ICC[2] world cup (takes place once in every four years among the national teams of all participating countries), IPL (Indian Premier League) and many more.

In limited over matches like ODI and T20, both sides bat once for a limited time (maximum 50 overs) with the aim in the first innings to score as many runs as possible, and in the second innings to score more than the target set in the first innings. Although ICC[2] officially have 20 nations participating in ODIs[1], more than 95% of all matches have been played by 9 main cricketing nations (Australia, England, India, Pakistan, West Indies, Sri Lanka, New Zealand, South Africa and Zimbabwe).

When some of the popular teams like India and Pakistan plays against each other, it is no longer considered a game between two teams but a game between two nations. Predicting the outcomes of such games is always of interest. Besides, Cricket is one such sport which holds an important place in the betting market. A high level of predictive ability with respect to outcomes is of great financial importance. Hence, we first present a formal problem statement.

2. PROBLEM STATEMENT

In this section, we elaborate upon the problem statement of predicting team standing in ODI for a Cricket League. To be able to do that, we are using some information as input, which we term as the unseen data. First, we consider all the participating teams in the league. The data set includes 20 teams but only 14 take part in the leagues. Second, Prior information of which team belongs to which group after they have been divided into 4 groups each having 4 teams. Rest all the information has been derived from the data of previously played matches already available to us.

We have used match information gathered from all 1005 ODI matches played after November 2006 till March 2015. We introduced a range of variables that could independently explain statistical significant proportions of variation associated with the predicted match outcomes. Such variables include home ground advantage, effect of playing the first innings, past performances, performance at the specific venue, performance against the specific opposition, experience at the specific venue and a few more.

Having results from these independent features, we combined the outcomes from these features to come up with predicting the winner for the game played between team A and team B.

This project report is organized as follows: Section 3 discusses the relevant prior work. Section 4 highlights different statistics of the dataset and its description. In Section 5, we present– the identification of features affecting quality and initial findings. Sections 6 and 7 respectively talk about the required mathematical background and algorithms which will be used, and Section 8 points the difficulties in the problem statement.

Our project is oriented toward a prediction using several machine learning algorithms that gives an estimate about the result of a match. Our algorithm uses past results to make a prediction about the future.

To be specific we are interested in predicting the team standings for the world cup 2015 before the world cup begins. We wish to predict the overall winner as well as the winner in each tier of the league. This can be visualized as a bottom-up tree, in which each team starts at the bottom and struggles to find a way up till the top.

3. RELEVANT PRIOR WORK

In this section, we briefly highlight the various approaches that have been chosen by various researchers who have worked on predicting the outcome of a game using several predictive techniques.

Bruce Morley & Dennis Thomas *et. al.* [6] examined the factors affecting the outcome of cricket matches played in the English one-day county cricket league. In particular, they focused on the home-field effect and the importance of winning the toss.

[1] **ODI** – One Day International played between two teams where both teams get 50 overs (maximum) each to bat and ball.

[2] **ICC –** International Cricket Council is the international governing body of cricket.

P. E. Allsopp and Stephen R. Clarke *et. al*. [3] applied techniques to determine the relative batting and bowling strengths and a common home advantage for teams playing both innings of international one-day cricket and the first innings of a test-match.

As seen in the above paragraphs, each of the researchers have attempted to predict the outcome of the game. For out predictive analysis we will follow a similar approach. As described in Section 1 and 2, our final goal of this project is to predict the team standings after a league has finished. The sub problem here is to be able to predict correctly the outcome of a game using several factors, including the past performance of a team.

4. DATASET DESCRIPTION AND STATISTICS

We found the cricket data set on cricksheet.org [1]. The data set that we have chosen comes with detailed information of 1077 One Day Internationals (ODI) played between various teams between November 2006 and March 2015. For every ODI match, one separate file has been logged. All these files are in YAML format. Every file can be broadly partitioned into two main parts, 1) overall summary and 2) Ball by ball information.

Though the size of the data set was not very big but the data being present in YAML format was difficult to parse. We wrote a script Python in order to access data available at all levels in the hierarchy. While most of the files followed similar structure, some files did not include a few information fields. For instance, if a match had no results due to rain or any other reason, the file would skip the result tag. There could be other fields missing which one can't know in advance until the file is opened and searched for them individually. So we decided to improve our script for parsing and this time included all possible corner cases while parsing the data file by file.

Table 1: Match counts, and win percentage of top 9 contributing teams

Team Name	# Matches		% Win
	Played	Won	
Australia	227	145	63.86
South Africa	180	111	61.66
India	254	149	58.66
Sri Lanka	238	122	51.26
Pakistan	201	101	50.24
England	204	96	47.05
New Zealand	183	86	46.99
Bangladesh	152	61	40.13
West Indies	166	57	34.33
Zimbabwe	125	28	22.4

Table 1, gives brief statistics of team's total matches played and won. As mentioned earlier, there are 20 teams that participates in the league but here are the top 9 teams which constitutes more that 95% of the total matches played. A thing to notice here is that the #Matches Played exceed the count of 1077 already. This is because a match between say India Vs Pakistan is one match but is counted individually for India and Pakistan.

The second part of each file gives ball by ball information. It mentions who is the player on Strike and Non-Strike, who is the bowler, how many runs are scored in that ball and wicket fall if any.

5. IDENTIFICATION OF FEATURES, INITIAL WORK AND FINDINGS

In this section, we present a detailed overview of the activities that have been done so far, and showcases the results of the experiments that we have performed. Besides this, we also present a sub-section to explain which novel features we have included which have an impact on winning probability of the team. One of the major task in this project is *feature identification.* In our predictive analysis using the machine learning model, we need to identify features that are consistent with the statistical findings. After coming up with a list of relevant features we need to shortlist them based on the accuracy they bring into the model. We have discussed the algorithm that we use to shortlist the features in section 7. Let's discuss feature identification in section 5.1.

5.1 Feature Identification

The most important feature that we identified is the advantage of playing in the **Home Ground**. There are various intuitive and established facts that says why a team playing in front of the home crowd in a home ground has more chances of winning the game [2].

Table 2: Showing % win of teams for the matches played at the home ground (HG).

Team Name	# Matches at the HG		% Win
	Played	Won	
Australia	99	68	68.68
South Africa	70	46	65.71
India	88	58	65.90
Sri Lanka	92	51	55.43
England	94	50	53.19

Table 2 provides the evidence that home advantage exists is overwhelming. The home winning percentage deviates significantly from the assumption of no advantage (the null hypothesis; p=0.5) within all major teams using a simple Binomial test.

Apart from the intuition that Home Ground is one of the most important feature, we conduct a Pearson Chi squared test, also known as goodness-of-fit test, to confirm whether there's any kind of dependence between the Home Ground feature and the result of the match. We conduct this test for every team. We present the case of the Pakistan team as an example in table 6.1. We calculate the chi-squared statistic and find it to be greater than the critical value for the significance level of 5%. Hence, we conclude that there's some association between the Home Ground feature and the result of a match.

Next important feature is the difference in the **ranking of the teams**. The rankings are provided by ICC which are based on the past performance of the teams. Unfortunately, this data set does not include the ranking of the teams. Neither does ICC preserve the past rankings. Every quarter they publish a ranking and overwrite the past ranking. However, they have mentioned on their website the method they use to rank the teams.

We have used the same method. With the rankings in order, we can come up with pair of teams for which we are most certain that the higher ranked team will defeat the lower ranked team.

Team / Rank on basis of %win		**% matches won in head to head by team**	
A	**B**	**A**	**B**
Australia/1	West Indies/8	89	11
India/3	South Africa/2	56	46

Table 3: Finding pairs of team that when plays against each other have better predictability of winner

As it can be noticed from the entries in above table, pair of teams close in ranking (like India/3 and South Africa/2) will not have a lot of difference in the % win when they are playing against each other. This can't be identified as the pair that does not give a clear indication of who the winner will be. On the other hand, if the teams are far apart (Australia and West Indies) in the ranking, they provide a higher level of certainty for the winner.

Some other features we are considering for finding out which team will be a winner are –

<u>All Features (including the two discussed above) considered while predicting the outcome:</u>

1. ***Advantage of playing in the Home Ground****:* Team playing in home ground has been observed to have won more matches than played away from home.
2. ***Ranking of the teams****:* Teams are ranked according to their past performances. Difference between these rankings can give us certain useful insights of who will be the winner. The higher the difference, the most probable is the result.
3. ***Team Batting First:*** Team that gets to bat first tries to score as many runs as possible. This gives the team an advantage of playing with a better
4. ***Teams average scores****:* What is the average runs scored by a team. This can give a comparison between the two teams on the basis of their average scores.
5. ***Performance at the specific venue:*** We have already seen how home ground plays an important role in team's chances of winning. Here by location we mean, given the location of the match happening, what are the chances of the team winning depending on the past records of the individual teams winning on that ground.
6. ***Performance against the specific opposition:*** How many times have the team won against a particular opposition in head to head match.

5.2 Initial Work and Findings

At the end of the match, summary of the match has been created. We have written a script in python that can generate this graph given below by just passing the file as input. Given below is the graph that shows the runs scored by both the teams as the game advanced.

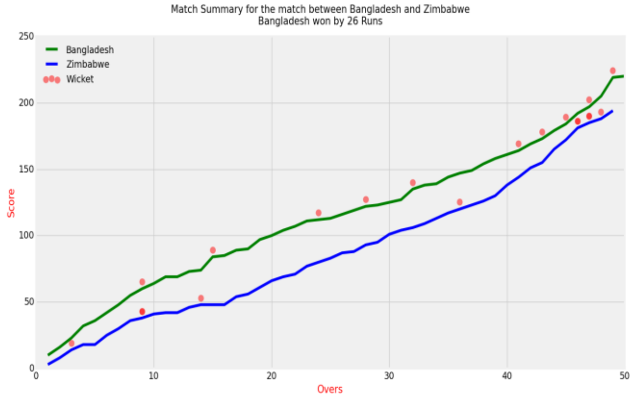

Figure 5.1: Match Summary on Score total for Bangladesh Vs Zimbabwe along with fall of wickets.

Figure 5.1 shows the progress in terms of runs scored and fall of wickets (represented by the red dots) over the span of 50 overs.

We found a few national teams to be very actively involved in playing ODI that the others. The Chord Diagram below shows this with color intensity varying as per the number of matches played between those two teams.

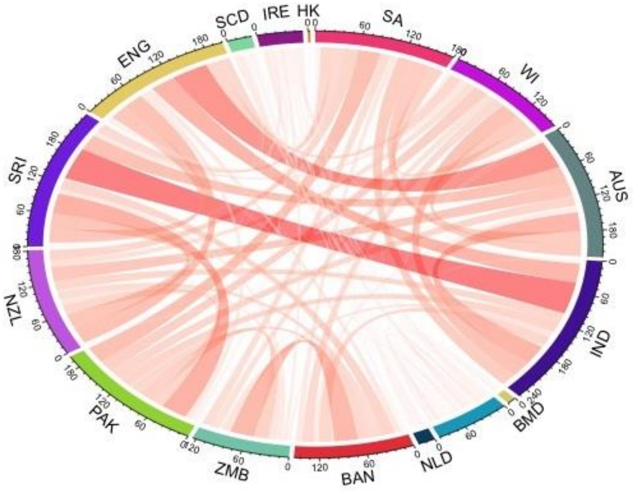

Figure 5.2: Chord Diagram showing number of matches played between the national teams.

Figure 5.2 justifies our initial finding that out of the 20 national teams playing cricket, 9 teams are more active that the others and accounts for 95% of the matches played. The graph has a chord between two teams if they have played against each other. The color intensity and width of the chord is directly proportional to the number of matches played between the two teams. For instance, more matches have been played between Sri Lanka & India and England & Australia.

We also come with another interesting finding that which's a stacked column plot for all the teams. We are interested in evaluating the performance of teams by comparing the fraction of matches it has won/lost. Figure 5.3 summarizes these findings.

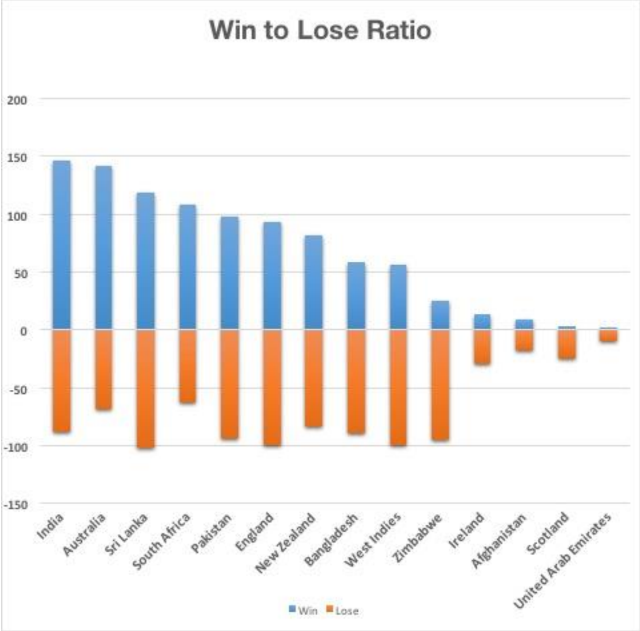

Figure 5.3: Stacked Column Plot

6. REQUIRED MATHEMATICAL BACKGROUND

Mathematical background required for regression techniques, Binomial Test, Hypothesis testing, ANOVA, Fischer's Test, Pearson Chi- squared test, and are discussed as follows:

Binomial test: The binomial test is an exact test to compare the observed distribution to the expected distribution when there are only two categories Our findings are the results of having done 'n' experiments, or having made 'n' observations, or having studied a sample of size 'n'.

t-Test: A t-test compares the mean of two samples. For instance, we have two samples of students of grade 10. First group of students are from China, while second belong to USA. We wish to find that whether there's any significant difference between any of the feature, say, SAT score, for the two groups.

ANOVA: When we have two samples as above, we construct a t-test. However, when we have more than two samples we use the ANOVA test. We use Fischer's Test to compute the result. If μ_0, μ_1, μ_2, . . . , μ_k are the means of k different samples. Then out null and alternative hypothesis that we wish to evaluate are:

$$H_0: \mu_0 = \mu_1 = \mu_2 = . . . = \mu_k$$

H_1: Means are not equal.

Chi-Squared Test: The test is applied when we have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables. The motivation to test independence of variables in our project is that we want to seek the factors that determine whether a team will win a particular match. We shortlist several factors that have been discussed in section 5.

To test the feature involving Home Ground Advantage, we chose a sample. We categorized all the matches played by the team Pakistan into following categories- Home Ground and Result.

	Home	Away	Total
Win	21	80	101
Loss	8	86	94
No Result	1	5	6
Total	30	171	201

Table 6.1: Contingency Table for Pearson Chi Squared Test

We wish to determine whether there is any significant association between the two variables Home Ground and Result. We apply the Pearson chi- Squared test and obtain the following results:

$$\chi^2 = 5.798 \text{ with 2 degrees of freedom}$$

Whereas, χ^2= 4.605 for 2 degrees of freedom for α=0.1 and for χ^2= 5.991 for 2 degrees of freedom for α = 0.05. Thus the variable Home Ground plays a significant role in determining a match's result. In a similar way we find other such factors that influence the outcome of a match.

Linear regression: A linear regression defines a relation between a single dependent variable and an outcome.

Multiple regression: A multiple regression extends this relation to more than one variable that determine the output.

The structural model for multiple regression is presented as:

$$E(Y|u, v, w) = \beta_0 + \beta_1 u + \beta_2 v + \beta_3 w$$

Here we can model the *'Y'* as team's winning probability. The input variables *u, v, w* are the dependent variable on which the outcome *Y* is dependent.

Bayes Classification Rule: The classification can be performed using Bayes Classification rule [13]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

7. ALGORITHMS AND METHODS

Here we discuss the division of data into training and testing dataset, the algorithm used to decide the match winner of an individual game and then computing the final team standings.

7.1 Splitting into training and testing

we first divided the dataset into Training and testing as follows-
Training – 80% and Testing – 20%

We trained our model on the training data and test on the testing data. We will be using cross validation score to find out the best model.

7.2 Algorithm to select the best model

Now, we discuss the algorithm we have implemented in selecting the most appropriate features for our model. To construct the baseline model we start with two features. Apart from these two features we have 6 more features that we may or may not include into our model. Suppose we build our baseline model with features f_1 and f_2. Furthermore, suppose our another six features are g_1, g_2, g_3, g_4, $g_5$and g_6. Let F() be a function that evaluates the accuracy given by the features that it takes as arguments, then, one of the possible algorithm for the feature selection process is given by:

```
Int max= $F(f_1, f_2, g_1)$
For i=2 to 6:
        For j=i+1 to 6:
                If (max< $F(f_1, f_2, g_i, g_j)$)
                        max= $F(f_1, f_2, g_i, g_j)$
        ForEnd
ForEnd
return max;
```

We realize that the complexity of the above algorithm is $O(n^2)$, and since it is a greedy algorithm, it appears to be optimal; however, it skips many of the comparisons that would have otherwise been made. To make every possible comparison, we need to generate every possible permutation of the feature set which is 2^n possible combinations. Therefore, to improve the complexity of our algorithm we use the following approaches for our classification problem, instead of making a brute force search.

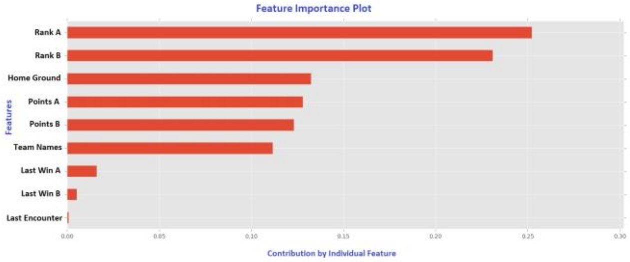

Table 7.1: Plot showing feature importance undertaken by classifier

We can model our problem as a Classification problem. We want the output in terms of which team is more likely to win the game based on some features. The output therefore is a categorical variable which depends on several other variables as discussed in section 5. These variables can be categorical like the location of the match, team winning more number of times in head to head game or it could be continuous like the ranking of the team.

Since we have both kind of variables, a decision based classifier and Regression Tree is best suited. It uses a vital concept of recursive partitioning to determine a split (decision) at each node.

Recursive Partitioning: In recursive partitioning the goal is to produce a node that is homogenous in nature and is distinct from other nodes. Hence it can represent a single class unambiguously [5]. As we progress from the root level to the next level, a splitting rule is applied to determine the path to be followed along the tree. This splitting should continue until either all leaf nodes contains inly observations of a single class or minimum number of observations in a single node of the leaf tree has been achieved. This address our main problem of finding the relative importance of the features which inherently determines which rule needs to be associated with which feature.

Classification: These are used for modeling data that have dependent variable as a combination of both categorical and continuous. Based on the recursive partitioning rules [4], the nodes of the decision tree are recursively created, and a structure that has all leaf nodes as distinct classes is identified. Now given the information about features like the number of groups and the teams in each group, the location for the match, we can provide these as inputs the prediction model which can determine which team out of the two playing against each other at every level (quarter final, semifinal and finals) will win the match. This way we will create a ranking/standing of the teams where the winner will be on top followed by the first runner up, second runner up and so on.

7.3 Finding the team Standings

The teams in world cup are divided into two pools, each having 7 teams. Teams in individual pool play against each team in the same pool. Every time the winner is awarded 2 points. In case of a draw, both teams earns 1 points each and looser gets nothing. So, in each pool, there will be 7 choose 2 = 21 matches. Similarly 21 matches in pool B. We maintain a table per pool with the points of each team. Top 4 teams from each pool advances to the next stage of qualifier. From here on, the matches are knockout matches.

There are 4 matches being played between these 8 teams at qualifier round and we get 4 winners. These 4 winner advance to semifinals, play a total of 2 matches and we get the finalist. The one that wins the final is the winner. Finally we build the team standings base on the points each team earned.

8. Limitation of the Model

As already defined, the goal of our project is to determine the outcome of a cricket match. While we approach to our goal using several predictive analysis techniques, we face the following difficulties and practical issues:

Past Performance of a Team: While this can be a good factor in our prediction, there are several issues that bother us. For instance, if a team has been in form throughout the current season and has therefore a significant chance of winning, however, few of the best performing players are injured and will not be playing the next game. In these cases, although, our predictive algorithm will favor the current team, the current team has very little chances of winning.

Duckworth Lewis: Unlike Soccer, a game of cricket is immediately paused during bad weather such as rain, storm etc. In such circumstances if weather is not restored shortly, then Duckworth Lewis method is applied which basically takes into account the performance of both the teams (if first team has finished its innings) and predicts that which tem would win in the current circumstances. There has been a lot of debate regarding the accountability of this method since it takes into account the performance of the team for the complete innings and not just the trend from the last few overs (something like Time Series analysis). Therefore, in such circumstances, our predictive technique may not be consistent.

Dynamic Team Selection: Contrary to the above issue of 'Past Performance of Team', if there were a scenario in which few of the players who were not in form are replaced by the promising ones then it's again a huge challenge for us.

9. FUTURE RESEARCH

In this research we have predicted the standings of team using the fact that each team is a single entity. We did not consider team as a combination of players, but rather, as a collective team. In future we'd like to work upon the prediction using teams as a selection of several players. Since we have ball by ball data given in which we know which player plays in excellent form against which player, we can create a model in which a team's performance would depend upon the performance of its players.

This is a pertinent feature in cricket since we often hear about players getting injured before a particular match, thereby, unexpectedly affecting the team's performance drastically.

10. CONCLUSION

In this project report, we have presented the entire set of activities that we have undertaken as a part of the project. After we selected the best model using Random Forest classifier, we tested out model on the test data and found these scores-

Training Score	0.6468
Testing Score	0.6363
Cross Validation Score	0.6281

Our model was able to predict 33 out of 49 games in ICC world cup 2015 thus predicting 67% of the total matches correctly.

The team standing in the end were calculated on the basis of who is winning in the every knockout matches and the teams that could not advance to knockout matches were ranked according to the points they earned.

11. REFERENCES

[1] Cricket data set downloaded from: http://cricsheet.org/

[2] *Alan M. Nevill , Roger L. Holder* : An Overview of Studies on the Advantage of Playing at Home. http://link.springer.com/article/10.2165/00007256-199928040-00001

[3] *P. E. Allsopp and Stephen R. Clarke*: Rating teams and analyzing outcomes in one-day and test cricket. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2004.00505.x/full

[4] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees.* New York: Chapman and Hall; 1984

[5] Strobl C., Malley J., Tutz G., *An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests.*

[6] *Bruce Morley & Dennis Thomas:* An investigation of home advantage and other factors affecting outcomes in English one-day cricket matches. http://www.tandfonline.com/doi/abs/10.1080/02640410410001730133

[7] *David Forresta & Ron Dorseyb*: Effect of toss and weather on County Cricket Championship outcomes. http://www.tandfonline.com/doi/abs/10.1080/02640410701287271

[8] Paper on Multinomial Logistic Regression: http://personal.lse.ac.uk/gerber/Slides/MY452L_2012_SeminarW8.pdf

[9] Blog on prediction using R: http://www.r-bloggers.com/introducing-cricketr-an-r-package-to-analyze-performances-of-cricketers/

[10] Parsing in Python: http://stackoverflow.com/questions/1773805/how-can-i-parse-a-yaml-file-in-python

[11] PyYaml Documentation: http://pyyaml.org/wiki/PyYAMLDocumentation

[12] Sample Problem statement on Cricket dataset: http://www.kenbenoit.net/courses/ME104/Lab9_ME104.pdf

[13] Machine Learning Predictions of Playoff Series: http://nhlnumbers.com/2013/9/6/machine-learning-predictions-of-playoff-series

[14] The anomalous contraction of the Duckworth-Lewis method: http://www.espncricinfo.com/magazine/content/story/459431.html

[15] Predicting cricket scores using linear regression: https://www.reddit.com/r/MachineLearning/comments/1ma2t4/predicting_cricket_scores_using_linear_regression

[16] India Vs Pakistan - Blog: The Analytics Compass http://www.simafore.com/blog/topic/cricket-statistics