Statistical and Predictive Analysis of Team and Player Performance in the Indian Premier League.

Damien Breen

T00171168

Code: https://github.com/DB79/BigData\_IPL\_Analysis

Abstract

This paper aims to determine if statistical and predictive analysis can be used to improve team selection in the Indian Premier League, similar to idea put forward in the Moneyball hypothesis. A number of key batting and bowler factors were identified and analyzed to determine if correlations could be found between different aspects of team and player performance. These correlations were then used as the basis for the predictive analysis which looked to predict the outcome of a delivery and in turn the best player to select under certain conditions.

# Introduction

The following report will look to analysis the performance of players and teams in the Indian Premier League. The Indian Premier League is a T20 format tournament and was founded by the Board of Control for Cricket in India in 2007. The league is contested every year between April and May by franchise teams from various Indian cities. Currently eight teams compete in the tournament and teams play each other twice in a home and away league format. At the end of this phase the top 4 teams advance to the playoffs. In the playoffs the top 2 teams play each other with the winner advancing straight to the final, while the loser will get a second chance and will play the winner of the match between the 3rd and 4th placed teams to determine the second finalist.

As the league format is T20 each teams has 20 overs with each over comprising of 6 deliveries. The game is described as having two innings, the first innings is the first set of 20 overs for one team while the roles (batting/bowing) are reversed in the second innings.

Teams are formed through competitive bidding from a collection of local as well as foreign players. Each team can purchase a maximum of 10 overseas players, out which only four can be members of playing eleven for any given match (Dar, 2016)*.*

The idea for this report was based on the Moneyball hypothesis put forward by (Lewis, 2003) and subsequent research by (Hakes & Sauer, 2006) which suggests baseball teams can improve results by taking a statistical approach when acquiring players. This in turn can led to improved results with a limited financial output.

# Approach

The data which will be used in this report contains two different datasets. The first dataset contains information about all of the Indian Premier League games which have been played since its inception in 2007 until the end of the 2016 season. Information such as toss winner (team), toss decision (bat or bowl), winning margin, venue and umpires can be found in this dataset.

The other dataset contains information about every delivery bowled during each of these games. This includes the batsman’s name, bowler’s name, total runs, a breakdown of how total runs are made up and information about the wicket, if a wicket was taken.

The first step for this paper was to remove some of the observations from the datasets. In the initial dataset there were three potential outcomes for each match, ‘normal’, ‘tie’ or ‘no result’. A match with a ‘normal’ result was kept in the dataset as one of the two teams won the match. A match with ‘tie’ as the result was also kept as the match was tied but a super over was used to determine the winner. Any match with ‘no result’, as the result type, was removed as the match could not be played typically because of the weather. In total this preprocessing only removed 3 observations but it was necessary as these observations would have impacted on the team performance analysis which will be carried out.

The dl\_applied variable was also used to remove observations from the two datasets. In cricket when weather affects play and teams are unable to complete their innings the Duckworth Lewis Method (DL) is used to determine the outcome. The Duckworth–Lewis method (D/L method) is a mathematical formulation designed to calculate the target score for the team batting second in a limited overs match interrupted by weather or other circumstances (Mankad, et al., 2014). All observations where the D/L method was used to determine the outcome were removed as a players performance would be affected and thus would have impacted the results of the analysis carried out in this paper. In total this resulted in 2,845 observations being removed from the deliveries dataset and 15 matches from the matches from the dataset.

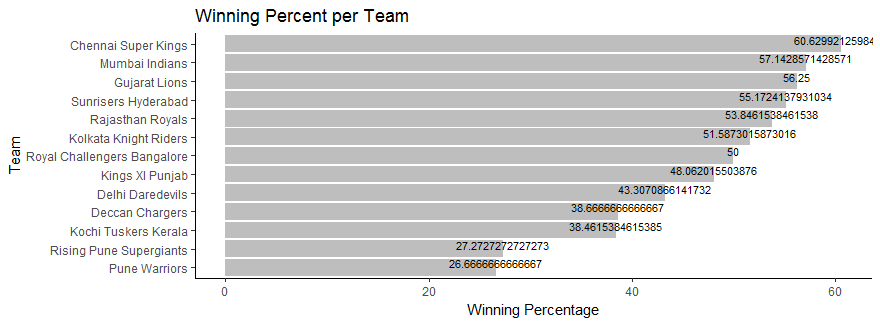
# Discussion

The game of cricket is one which hinges on a number of key skills, including batting, bowling and fielding. In this paper the focus is going to be on bowling and batting as these skills are more measurable than fielding which would be difficult to analysis statistically.

The batting and bowling performance of both teams and individual players will be analyzed to determine if a pattern can be identified and then the findings will be used as the basis for the predicative analysis which will be conducted.

## Team Analysis

Firstly some analysis was done of the matches’ dataset to determine the win percentage of each of the thirteen teams who have completed in the IPL. Percentage was used, instead of number of wins, as not all teams have played a comparative number of matches and thus percentage will give a better representation of the data. The results of this analysis can be seen in the *Figure 1* below.

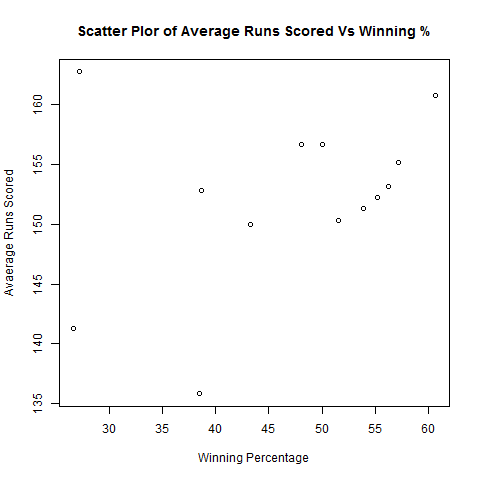


*Figure 1 - Winning Percentage per Team*

From this analysis we can see that the team with the highest win percentage is the Chennai Super Kings (60.63%) while Pune Warriors (26.67%) have the lowest.

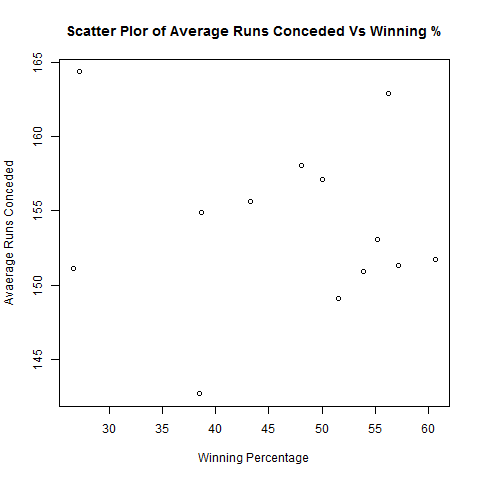
The next analysis involved calculating the average scores per match per team. This was done to see if a teams’ chances of winning related to the number of runs they score in a match. The results can be seen in *Figure 2*.

The results here show that there is a weak correlation (0.32688) between runs scored and winning percentage.



*Figure 2- Scatter Plot of Average Runs Scored Vs Teams Winning %*

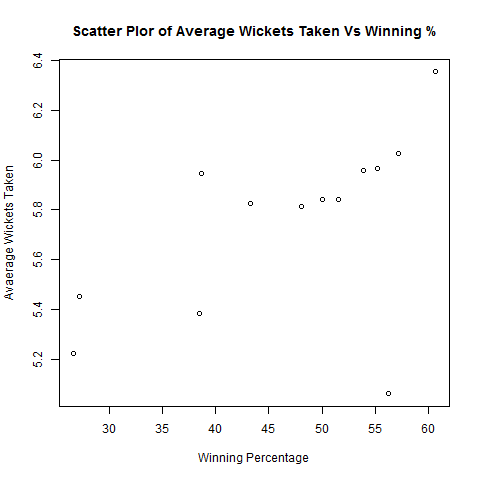
The next analysis will try to determine if average runs conceded is a better indicator of a team chances of winning a match. The results of the analysis can be seen in *Figure 3*.



*Figure 3 - Scatter Plot of Average Runs Conceded Vs Winning %*

The correlation here was much weaker (-0.0983) and this shows that the outcome of a match can’t be determined by the number of runs conceded.

The next factor that will be considered is the number of wickets taken by competing teams to see if this can be used to predict the outcome of a match. The results of the analysis can be seen in *Figure 4*.



*Figure 4 - Scatter Plot of Average No. of Wickets Taken Vs Winning %*

From this analysis we can see that there is a stronger correlation (0.54822) between a teams’ chances of winning and the number of wickets taken by a team.

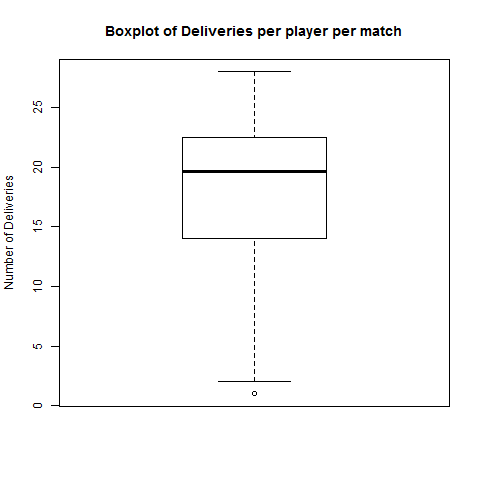
From this analysis, based on winning percentage, it can be concluded that runs scored and conceded can’t be used to determine the outcome of a match while the number of wickets taken proved to be a better indicator of the outcome.

All of the analysis thus far has looked at team performance and the next focus will be on individual player performance with the focus again on batting and bowling performance.

## Player Analysis

To be able to perform analysis of player performance some analysis was carried out to calculate player averages for both bowlers and batsmen. For bowlers, average number of deliveries bowled, runs conceded, dot balls and wickets taken were calculated. For batsmen, averages were calculated for balls faced, runs scored, boundaries scored and strike rate.

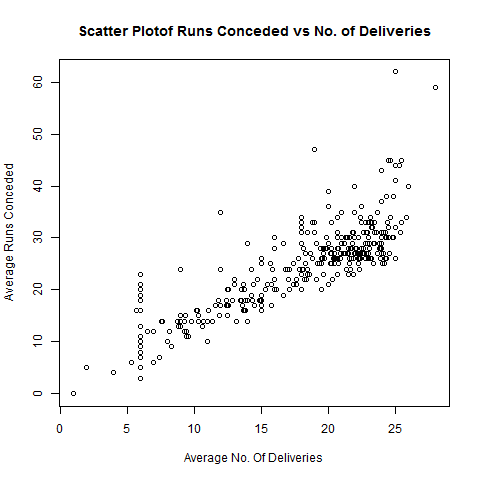
In *Figure 5* a boxplot is used to display the average number of balls bowled per player per match.



*Figure 5 - Boxplot of Average Number of Deliveries per Player per Match*

From this boxplot the mean number of deliveries is 17.98 and the median is 19.68. These stats for deliveries will be used to determine if there is a correlation between number of deliveries and other bowling factors such as dot balls, wickets and runs conceded.

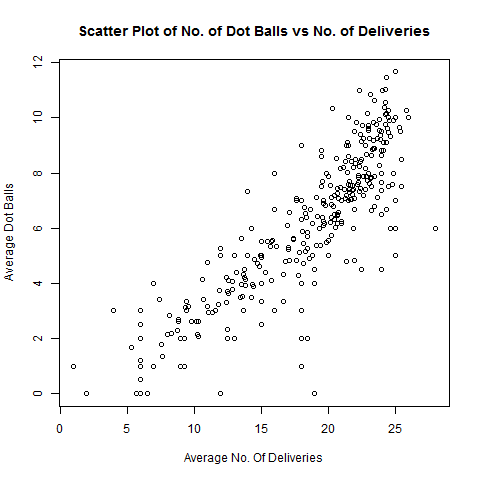
*Figure 6* shows the correlation between average number of runs conceded and number of deliveries bowled.



*Figure 6 - Scatter Plot of Run Conceded Vs No. of Deliveries*

From this scatter plot we can see that there is a strong correlation (0.8297) between the number of runs conceded and the number of deliveries bowled by a bowler. This was the expected result, as logically you would expect the number of runs conceded to increase with every delivery bowled.

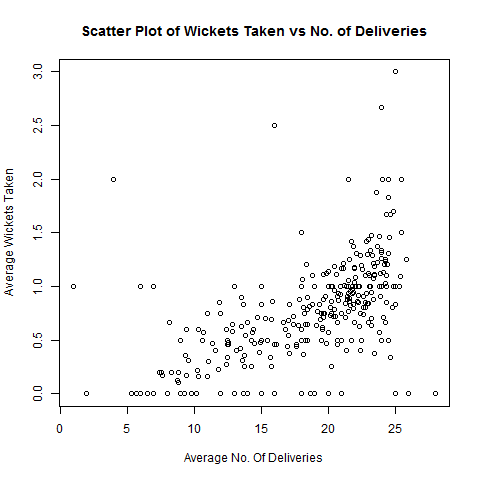
Similar analysis was carried out on dot balls and the results can be seen in *Figure 7*.



*Figure 7 – Scatter Plot of Dot Balls Vs No. of Deliveries*

There is again a strong correlation (0.86499) between dot balls and the number of deliveries and this was again the expected result.

The final bowling factor to be analyzed is correlation between wickets taken and the number of deliveries bowled by a bowler. The results can be seen in *Figure 8*.



*Figure 8 - Scatter Plot of Wickets Taken Vs No. of deliveries*

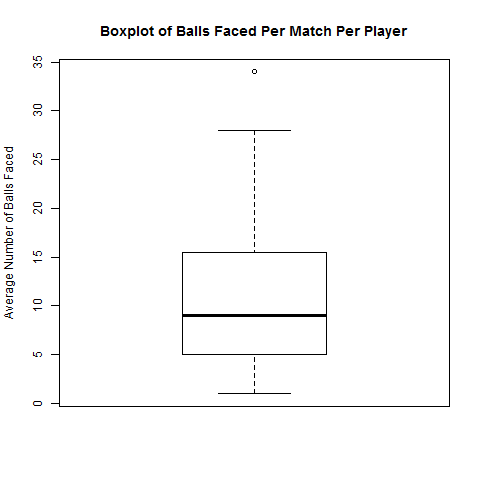
The correlation between wickets and the number of deliveries was relatively low, compared to the previously calculated correlations, at 0.6009. This was somewhat surprising as you would expect the number of wickets taken to increase as the number of balls bowled increases.

From the analysis carried out on bowling performance it can be concluded that the number of runs conceded and dot balls will increase as the number of deliveries increase. It was also proved to a lesser extent that the number of wickets taken will also increase with an increased number of deliveries.

Similar analysis was also performed in relation to batting performance to determine if the number of runs scored, boundaries scored (4 or 6 runs) and strike rate increased according to the number of balls faced. Strike Rate = 100 Runs/Balls, i.e. SR = the number of runs scored for every 100 balls faced (Lemmer, 2011).

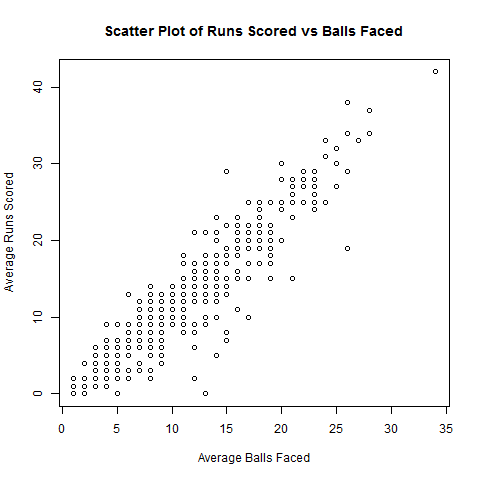
The boxplot in *Figure 9* shows the average number of balls faced per player per match.

This analysis found that the median number of balls faced by a batsman was 9.00 and the mean was 10.42. This data will be used to determine if there is a correlation between balls faced and other batting indicators such as runs scored, boundaries and strike rate.



*Figure 9 - Boxplot of Average Balls Faced per Player per Match*

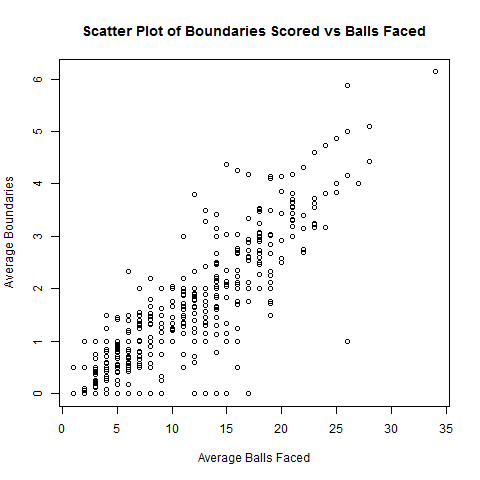
The correlation between runs scored and balls faced can be seen in *Figure 10.*



*Figure 10 - Scatter Plot of Runs Scored Vs Balls Faced*

From this scatter plot we can see that is a strong correlation (0.94099) between the number of runs scored and the number of balls faced by a batsman. This is as expected as logically the more balls you face the more runs you are likely to score.

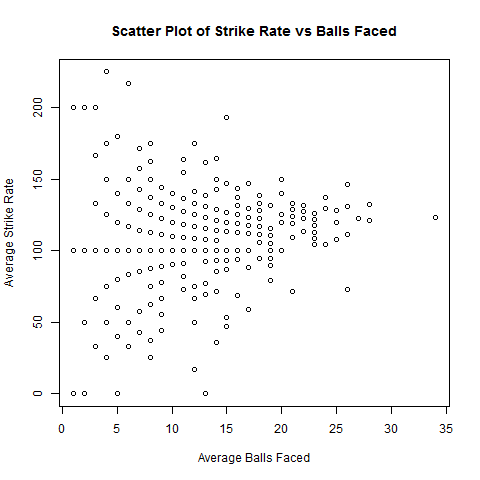
The correlation between boundaries and number of balls faced was also calculated and the results can be seen in Figure 11.



*Figure 11 - Scatter Plot of Boundaries Vs Balls Faced*

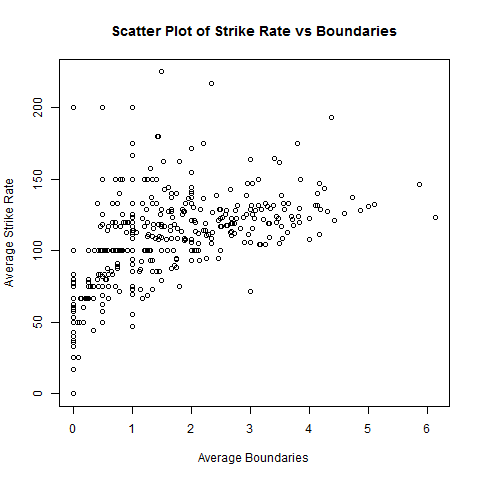
From this scatter plot we can see that there is again a strong correlation (0.8697) between the number of boundaries and the number of balls faced.

The next factor analyzed was a players’ strike rate and the results can be seen in *Figure 12*.



*Figure 12 - Scatter Plot of Strike Rate Vs Balls Faced*

The results of this analysis showed that there is a weak correlation (0.3456) between a players strike rate and the numbers of balls faced. This was somewhat expected and a better comparison would be between boundaries and strike rate. The results of this comparison can be seen in *Figure 13*.



*Figure 13 - Scatter Plot of Strike Rate vs Boundaries*

From this analysis we can see that there is a relatively stronger correlation (0.5708) between a players strike rate and the number of boundaries they have. This makes sense as by scoring a boundary (4 or 6 runs) a batsman is significantly increasing their average runs per ball. This is different to the previous comparison as just because a player faces an increase number of balls there is no guarantee that their average runs per ball will increase and this is why there is a weak correlation (0.3456) between strike rate and balls faced.

From this analysis of player batting performance it was discovered that there is strong correlation between the number of balls faced and runs scored and between balls faced and boundaries. However there is no significant correlation between balls faced and a players’ strike rate. There is however, a stronger correlation between a players’ strike rate a boundaries.

## Predictive Analysis

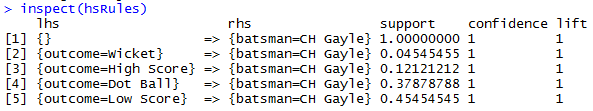
From the analysis conducted thus far we can see that when trying the predicate the outcome of a match the number of wickets taken provides the best correlation. Based on this we can say a team will have a better chance of winning if they increase their chance of taking wickets.

Association rules (Apriori algorithm) was used to determine if the outcome of a delivery can be predicted and by comparing the predictions for different players and is it possible to select a player with a higher chance of taking a wicket.

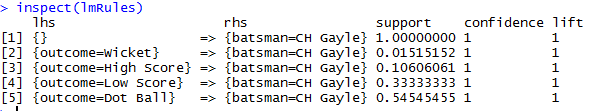
If we take an example where Chris Gayle is the batsman and he is playing against the Mumbai Indians. By filtering the data to include only observations where Gayle is the batsman and the bowlers are players available to Mumbai Indians we can try and determine if one player has a better chance of getting a wicket compared to the other.

If we select two Mumbai Indian bowlers, Harbhajan Singh and Lasith Malinga, the results can be used to determine who should bowl when Gayle is batting.

The results for both players can be seen below.



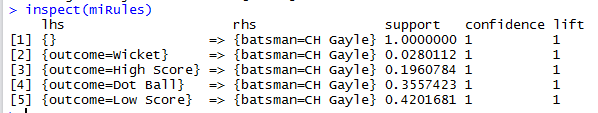
*Figure 15 - Gayle Vs Singh*



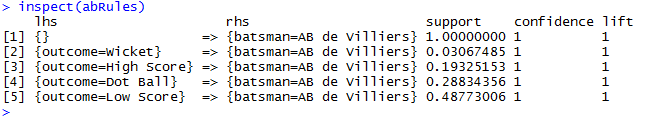
*Figure 14 - Gayle Vs Malinga*

From this we can see that Singh has a better chance of taking a wicket when compared the Malinga. Malinga only takes a wicket in 0.015 cases while Singh will take a wicket in 0.045 instances. The difference here is quite small and this shows how difficult it is to take a wicket but by selecting Singh to bowl against Gayle the Mumbai Indians can increase their chances of getting a wicket which in turn will increase their chance of winning based on the research carried out previously.

Similar analysis was also carried out to determine if a batsman’s performance could be predicted based on the opposition. Again Gayle was selected as of the batman with AB de Villiers also being selected and the opposing team chosen were the Mumbai Indians. The results of this analysis can be seen in Figures 16 and 17.



*Figure 16 - Gayle Vs Mumbai Indians*



*Figure 15 - AB de Villiers Vs Mumbai Indians*

From this we can see that when it comes to matches against Mumbai Indian there are some differences in terms of the players’ predicted performance. There is minimal difference when it comes to chances a wicket being taken (0.028 vs 0.030) and high score (0.196 vs 0.192). However when it comes to the other possible outcomes the results are more noticeably different. Gayle has 0.36 chance of getting a dot ball (no run scored) while de Villiers has a 0.29 chance of not scoring. This is significant as it means that Gayle is likely to score no runs with 36% of the balls he faces while this figure is 29% for de Villiers. Based on this one would say that de Villiers would be a better player to pick because he is likely to score runs with more of the balls compared to Gayle.

# Findings

In this paper it was found that it is quite difficult to predict the outcome of an IPL cricket match based on the number of runs scored, the number of runs conceded or the number of wickets taken by a team. The number of wickets taken proved to have the biggest impact on the outcome of the match.

It was also discovered that the number of deliveries by a bowler has a significant impact on the number of dot balls bowled and the number of runs conceded. The relationship between wickets taken and deliveries proved to be less significant.

In terms of batting, the results were quite similar as the number of balls faced has a strong correlation between the number of runs scored and boundaries scored but there was a weak correlation between balls faced and a players’ strike rate.

It was also discovered that given a choice of two bowlers, association rules can be used to select the bowler who has a better chance of taking a wicket. While the same can be applied to a choice of two batsmen against an opposing team.

# Conclusion

From the analysis carried out in this paper it can be concluded that statistical and predictive analysis can be used to improve team selection in the Indian Premier League. While the results showed minimal differences in players performance there was enough of a difference to suggest that the outcome would differ depending on which player was selected.

The results show that the Moneyball approach is something which showed be researched further with regard to it being used to improve team selection and results by teams in the Indian Premier League.

# References

Dar, U. R., 2016. Indian Premier League- Boon or Bane. *IOSR Journal of Sports and Physical Education,* 3(6), pp. 1-4.

Hakes, J. . K. & Sauer, R. . D., 2006. An Economic Evaluation of the Moneyball Hypothesis. *Journal of Economic Perspectives,* 20(3), pp. 173-185.

Lemmer, H. H., 2011. The Single Match Approach to Strike Rate Adjustments in Batting Performance Measures in Cricket. *Journal of Sports Science and Medicine,* 10(4), pp. 630-634.

Lewis, M., 2003. *Moneyball: The Art of Winning an Unfair Game.* New York: W. W. Norton & Company.

Mankad, S. . H., Chaudhary, A., Dalsaniya, N. & Mandir, V., 2014. Study and Analysis of Duckworth Lewis Method. *International Journal for Scientific Research & Development,* 2(2), pp. 469-473.