Statistical Analysis of IPL Player Performance using Advanced Computational Methods

Nirat J. Patel
L.D. College of Engineering
Ahmedabad

ABSTRACT

Cricket in India is rapidly expanding, largely due to the Indian Premier League (IPL), which is now valued at nearly $16.4 billion. This growth has led to the emergence of new opportunities in Sports Analytics in Cricket. The IPL is a franchise cricket tournament where teams acquire players through an auction format. With numerous choices available during the auction and a limited budget, it is crucial for franchises to make informed player selections and create well-balanced teams with strong batting, bowling, and fielding capabilities. In this paper, player performance was analyzed using the IPL matches and ball-by-ball dataset from 2008 to 2024, obtained from Kaggle. Batting performance was evaluated using adjusted batting scores and the stochastic dominance method for comparison, while bowling performance was assessed using a measure called Combined Bowling Rate (CBR), which combines three existing statistics. Python libraries such as Numpy for scientific computing, pandas for data analysis, and Matplotlib for visualization of the results were utilized for the analysis.

Keywords

Data Analysis, Indian Premier League, Player Performance, Cricket Analysis

1. INTRODUCTION

The field of sports analytics has gained significant traction over the past decade, revolutionizing how teams and analysts evaluate player performances and strategize for the games. Within cricket, the Indian Premier League (IPL) has emerged as one of the most prominent T20 Leagues globally. Given the high stakes and fast pace of the T20 game, it is of immense importance that the player's performance is measured accurately to optimize their strategies. There is also a prospect for fans who wish to gain deeper insights into the game’s beauty.

Traditionally in cricket, the player's performance has been evaluated using basic statistics such as batting average, batting strike-rate etc for batsmen and bowling average, bowling strike-rate and economy rate for bowlers. However, these metrics have certain assumptions that fail to capture the nuances of the game and thus might not be the most accurate metric. For e.g. the batting average metric does not consider the innings in which the batsman has remained not out. Thus, if a batsman remains not out very often in his career then an accurate measure of his batting performance may not be possible. Recognizing this, researchers have developed more sophisticated methods to assess player performance. Notable works include Damodaran (2008) and Lemmer (2002), who introduced statistical methods to evaluate batsmen and bowlers in One Day Internationals (ODIs). These methods provide a more comprehensive assessment by incorporating various factors.

Despite these advancements, there remains a gap in applying such metrics to the T20 format, particularly in the context of IPL. This paper aims to bridge this gap by implementing the established statistical methods from ODIs to T20 format, using the IPL as a case study. This paper translates the mathematical equations and functions from the aforementioned studies into Python thus creating a robust framework for analysing IPL player performance.

2. EXISTING SYSTEM

The field of cricket analytics has seen a significant rise over the years, with researchers developing various methods to evaluate player performances.

2.1 Batting Average Defined by Wood

In 1945, Wood [1] introduced the first measure for assessing batting performance by calculating the batting average of batsmen as the ratio of runs scored to the total number of innings played.

$$B_i = \frac{R_i}{n} \qquad (1)$$

Wood's metric considers not-out innings as completed, which can introduce biases depending on how the not-out innings are counted in the denominator.

2.2 Batting Performance Measure by Damodaran

Building on Wood's work, Damodaran [2] (2006) addressed the bias in calculating batting averages using a Bayesian approach and stochastic dominance to compare batsmen's performances. He argued that the traditional batting average fails to capture the various facets of a batsman's abilities. By assuming that a batsman's score follows a geometric distribution, where the chance of getting out is independent of the score, Damodaran estimated the number of runs a batsman would have scored in a not-out inning had they continued batting.

Assume that in his j^{th} innings player i remains ‘not-out’ and scored R_{ij}. Then a binary variable G_{rik} can be defined such that

$$G_{rik} = \begin{cases} 0 \; if \; R_{ik} < R_{ij} \\ 1 \; if \; R_{ik} \geq R_{ij} \end{cases} for \; k = 1, 2 \ldots, j-1 \qquad (2)$$

Define,

$$n_{ij} = \sum_{k=1}^{j-1} G_{rik} \qquad (3)$$

Define,

$$C_{ik} = \begin{cases} 0 \; if \; R_{ik} < R_{ij} \\ R_{ik} \; if \; R_{ik} \geq R_{ij} \end{cases} \quad for \; k = 1,2, \ldots, j-1 \quad (4)$$

The estimate number of runs that the ‘not out’ batsman would have gone on to score is then given by:

$$E_{ij} = \sum_{k=1}^{j-1} \frac{C_{ik}}{n_{ij}} \qquad (5)$$

In every instance the batsman remained not-out, the batsman's score in that innings is replaced by E_{ij}.

2.3 Bowling Measure by Lemmer.

When assessing a bowler's performance, it's crucial to consider three primary measures: bowling average, bowling strike rate, and economy rate. In 2002, Lemmer [3] introduced the combined bowling rate (CBR), which utilizes the harmonic mean to create a comprehensive evaluation metric encompassing all three statistics.

$$A = \frac{R}{W}$$

$$S = \frac{B}{W}$$

$$E = \frac{R}{O}$$

Where O = number of overs bowled by a bowler, B = number of bowls bowled by a bowler, R = Number of runs conceded, W = number of wickets taken by the bowler.

According to Kenny and Keeping [4], the harmonic mean is used to find the average of ratios such as rates. So, a combined metric for bowling performance can be defined as the harmonic mean of the metrics mentioned above as CBR (Combined Bowling Rate):

$$CBR = \frac{3}{\left(\frac{1}{A} + \frac{1}{S} + \frac{1}{E}\right)} \qquad (6)$$

3. METHODOLOGY

In the proposed system, a comprehensive approach is taken by leveraging Python libraries such as Pandas for in-depth data analysis, data cleaning, and feature engineering, as well as Numpy for complex scientific calculations. Furthermore, the utilization of Matplotlib for data visualization enhances the interpretability of the results. The process begins with the meticulous acquisition and cleaning of data, followed by the development of sophisticated algorithms in Python to precisely compute a diverse range of metrics, which, in turn, will be instrumental in the thorough analysis of player performance.

3.1 Data Collection

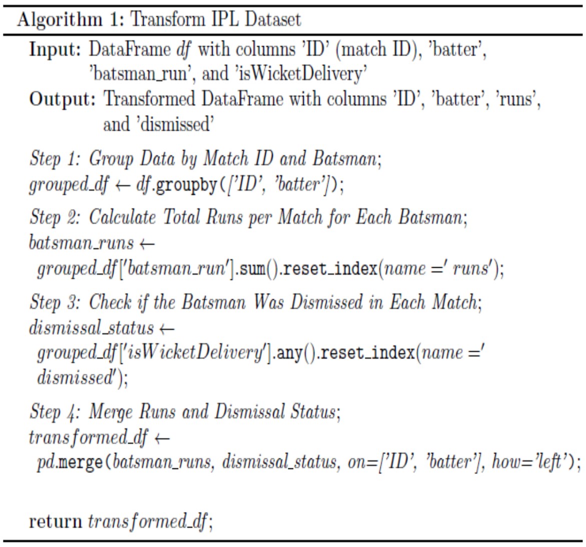
The datasets that have been used here are collected from Kaggle.com. 2 different datasets have been used for this paper. The first dataset is the matches data from Kaggle [5] containing the summary of each match for IPL 2008-2024. The matches are sorted from 2024 to 2008. The dataset contains 20 Columns containing data of 1095 matches. Another dataset that is used is also from Kaggle [6]. It contains data for each ball for the entire tournament from 2008-2024. It contains 17 columns and 260920 rows. A few columns that are present in the dataset are overs, bowl number, batter, bowler etc.

3.2 Data Preprocessing

The first step of data preprocessing is to clean the dataset for the null values. There are a few columns where almost all the rows have null values. So instead of dropping the rows, it is better to drop the columns, as they are of little importance to us. Also, there were team name changes for two teams since the season 2022, therefore the dataset contains two different names for the same team. Thus, the team names were made the same for both datasets. There are now 17 unique teams that have played IPL in total.

3.3 Metric Evaluation

With a clean dataset, the next step is to calculate the metrics for evaluating IPL players' performances. The evaluation and comparison of batsmen will be conducted using Damodaran's (2006) method. Damodaran noted that raw data, typically innings-by-innings scores, fail to account for not-out innings, where the score doesn't reflect the runs a batsman might have scored had they continued. To address this, Damodaran applied a Bayesian approach to estimate the likely runs for not-out innings. The following algorithm implements this approach:

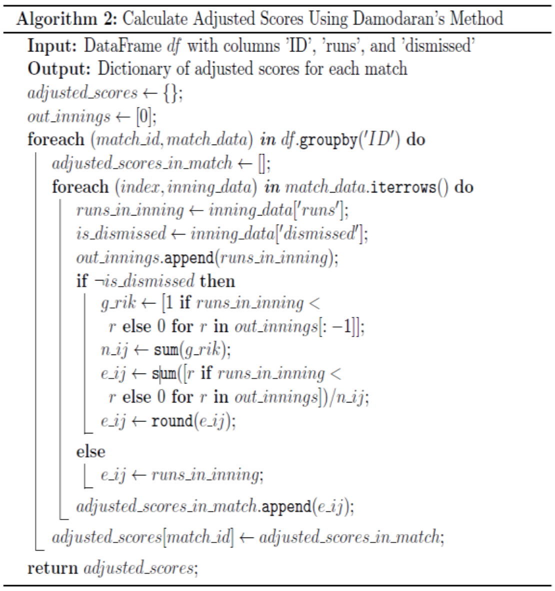


```
Algorithm 1: Transform IPL Dataset
  Input: DataFrame df with columns 'ID' (match ID), 'batter',
         'batsman_run', and 'isWicketDelivery'
  Output: Transformed DataFrame with columns 'ID', 'batter', 'runs',
          and 'dismissed'

  Step 1: Group Data by Match ID and Batsman;
  grouped_df ← df.groupby(['ID', 'batter']);
  Step 2: Calculate Total Runs per Match for Each Batsman;
  batsman_runs ←
    grouped_df['batsman_run'].sum().reset_index(name =' runs');
  Step 3: Check if the Batsman Was Dismissed in Each Match;
  dismissal_status ←
    grouped_df['isWicketDelivery'].any().reset_index(name ='
    dismissed');
  Step 4: Merge Runs and Dismissal Status;
  transformed_df ←
    pd.merge(batsman_runs, dismissal_status, on=['ID', 'batter'], how='left');

  return transformed_df;
```

First, all the rows belonging to a particular batsman of interest are filtered out. After that, the algorithm for calculating the adjusted scores can be applied as follows:



```
Algorithm 2: Calculate Adjusted Scores Using Damodaran's Method
  Input: DataFrame df with columns 'ID', 'runs', and 'dismissed'
  Output: Dictionary of adjusted scores for each match
  adjusted_scores ← {};
  out_innings ← [0];
  foreach (match_id, match_data) in df.groupby('ID') do
      adjusted_scores_in_match ← [];
      foreach (index, inning_data) in match_data.iterrows() do
          runs_in_inning ← inning_data['runs'];
          is_dismissed ← inning_data['dismissed'];
          out_innings.append(runs_in_inning);
          if ¬is_dismissed then
              g_rik ← [1 if runs_in_inning <
                r else 0 for r in out_innings[: −1]];
              n_ij ← sum(g_rik);
              e_ij ← sum([r if runs_in_inning <
                r else 0 for r in out_innings])/n_ij;
              e_ij ← round(e_ij);
          else
              e_ij ← runs_in_inning;
          adjusted_scores_in_match.append(e_ij);
      adjusted_scores[match_id] ← adjusted_scores_in_match;
  return adjusted_scores;
```

This metric is tested on two of the most prominent players of the IPL, Rohit Sharma and Virat Kohli. The datasets of these batsmen, along with their adjusted scores, are given in table 1 and 2:

Table 1: Replacing 'Not Out' scores with estimates of runs likely to have been scored by a batsman: Rohit Sharma's Randomly sampled 7 matches.

ID	Batter	Runs	Dismissed	Adjusted Score
980917	RG Sharma	7	True	7
1304105	RG Sharma	18	True	18
419117	RG Sharma	1	True	1
336007	RG Sharma	23	True	23
548308	RG Sharma	1	True	1
598037	RG Sharma	79	False	98
501246	RG Sharma	49	True	49

Table 2: Replacing 'Not Out' scores with estimates of runs likely to have been scored by a batsman: Virat Kohli's Randomly sampled 7 matches

ID	Batter	Runs	Dismissed	Adjusted Score
598057	V Kohli	17	True	17
1254058	V Kohli	33	True	33
548367	V Kohli	3	False	26
598054	V Kohli	99	True	99
733991	V Kohli	0	True	0
1254063	V Kohli	33	True	33
1136571	V Kohli	57	True	57

The adjusted scores data can be used to compare the performances of two batsmen using principles of stochastic dominance from finance. This method has recently been applied in sports analytics, such as in a paper comparing teams in the LaLiga football tournament by Fernández-Ponce, JM et al. [6]. The same approach will be employed to compare the performances of Rohit Sharma and Virat Kohli. Using the first-order stochastic dominance approach, it can be determined that a batsman A's performance is superior to that of another batsman B if, for any given score level, the probability of batsman A achieving a score greater than the given score is never lower, and sometimes higher, than the probability of batsman B achieving a score greater than that given score.

This can be done by plotting cumulative probability charts of the two batsmen with runs on the X axis and the probability of scoring more runs than the runs on the X axis value on the Y axis. Visually a batsman whose curve envelopes the other batsman's curve stochastically dominates the other batsman.

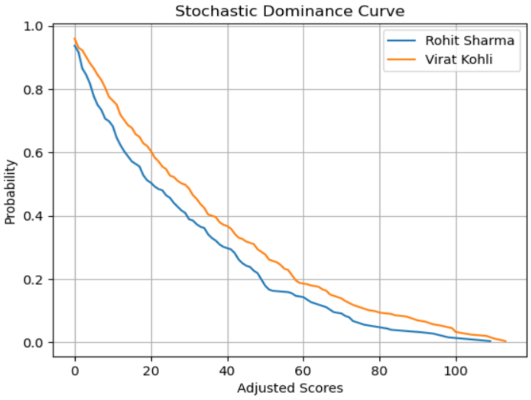


Figure 1: Stochastic dominance curves comparing Virat Kohli and Rohit Sharma

The stochastic dominance curve indicates that Virat Kohli consistently outperforms Rohit Sharma in scoring runs. Both players have played a similar number of matches in the IPL, making it a fair comparison. Another similar pair in terms of matches and playing style are Hardik Pandya and Shreyas Iyer.

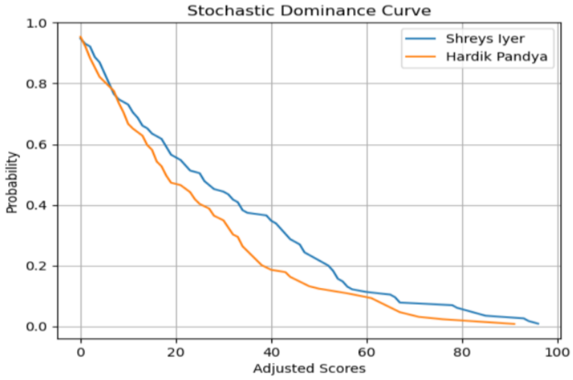
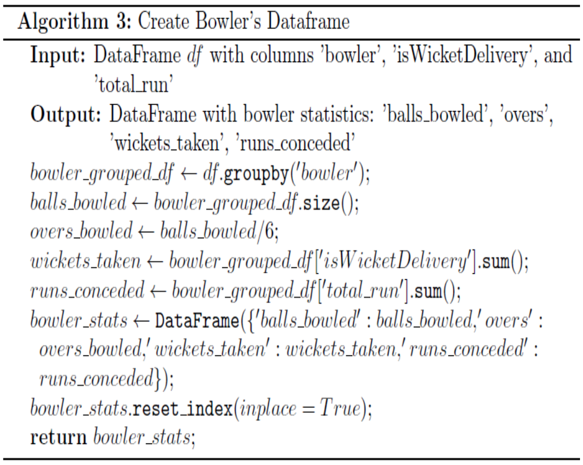


Figure 2: Stochastic dominance curve for Hardik Pandya and Shreyas Iyer

This describes the performance measure of batsmen. Next, the method for measuring and comparing the performances of bowlers in the IPL is considered. For this, the approach described by Lemmer in his 2002 paper will be used.

For the calculation of this metric, a dataset about the bowler's career information is needed. Below is an algorithm to create this dataset from ball-by-ball data



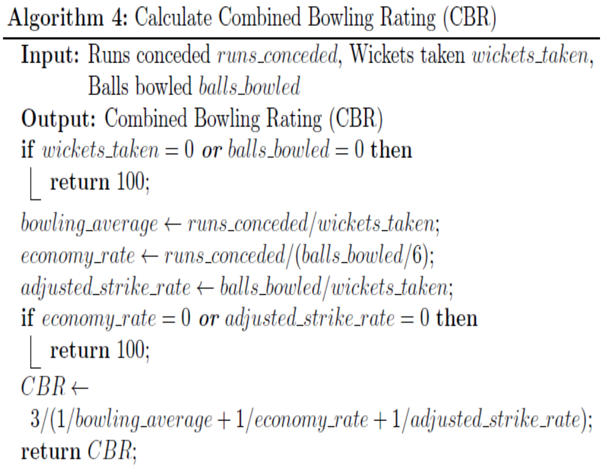
```
Algorithm 3: Create Bowler's Dataframe
  Input: DataFrame df with columns 'bowler', 'isWicketDelivery', and
         'total_run'
  Output: DataFrame with bowler statistics: 'balls_bowled', 'overs',
          'wickets_taken', 'runs_conceded'
  $bowler\_grouped\_df \leftarrow df.\texttt{groupby}('bowler')$;
  $balls\_bowled \leftarrow bowler\_grouped\_df.\texttt{size}()$;
  $overs\_bowled \leftarrow balls\_bowled/6$;
  $wickets\_taken \leftarrow bowler\_grouped\_df['isWicketDelivery'].\texttt{sum}()$;
  $runs\_conceded \leftarrow bowler\_grouped\_df['total\_run'].\texttt{sum}()$;
  $bowler\_stats \leftarrow \texttt{DataFrame}(\{'balls\_bowled' : balls\_bowled,' overs' :$
    $overs\_bowled,' wickets\_taken' : wickets\_taken,' runs\_conceded' :$
    $runs\_conceded\})$;
  $bowler\_stats.\texttt{reset\_index}(inplace = True)$;
  return $bowler\_stats$;
```

To ensure that the comparison of bowlers' performances is fair, the comparison is limited to only those bowlers who have bowled at least 100 overs in their IPL careers. This results in the following dataset: there are 125 bowlers in the IPL who have bowled at least 100 overs in their careers.

Table 3: Dataset of bowler's metrics

Bowler	Balls Bowled	Overs	Wickets Taken	Runs Conceded
SL MALINGA	2974	496	188	3486
RASHID KHAN	2901	484	157	3340
A KUMBLE	983	164	49	1089
IMRAN TAHIR	1340	223	86	1729
JJ BUMRAH	3185	531	182	3840
DP NANNES	689	115	38	815
MA STARC	879	146	59	1175
MM PATEL	1382	230	82	1733
MM ALI	770	128	40	900
SP NARINE	4146	691	200	4672

Now calculation of the CBR score for each bowler can be done using the following algorithm:



```
Algorithm 4: Calculate Combined Bowling Rating (CBR)
  Input: Runs conceded $runs\_conceded$, Wickets taken $wickets\_taken$,
         Balls bowled $balls\_bowled$
  Output: Combined Bowling Rating (CBR)
  if $wickets\_taken = 0$ or $balls\_bowled = 0$ then
      return 100;
  $bowling\_average \leftarrow runs\_conceded/wickets\_taken$;
  $economy\_rate \leftarrow runs\_conceded/(balls\_bowled/6)$;
  $adjusted\_strike\_rate \leftarrow balls\_bowled/wickets\_taken$;
  if $economy\_rate = 0$ or $adjusted\_strike\_rate = 0$ then
      return 100;
  $CBR \leftarrow$
    $3/(1/bowling\_average + 1/economy\_rate + 1/adjusted\_strike\_rate)$;
  return $CBR$;
```

Table 4 shows the top 10 bowlers of IPL based on their CBR values, the smaller the value the better the bowler can be considered.

Table 4: Top 10 Bowlers based on their CBR values

Bowler	Balls Bowled	Overs	Wickets Taken	Runs Conceded	CBR
SL MALINGA	2974	496	188	3486	11.568
RASHID KHAN	2901	484	157	3340	12.200
A KUMBLE	983	164	49	1089	12.230
IMRAN TAHIR	1340	223	86	1729	12.341
JJ BUMRAH	3185	531	182	3840	12.357
DP NANNES	689	115	38	815	12.362
MA STARC	879	146	59	1175	12.395
MM PATEL	1382	230	82	1733	12.522
MM ALI	770	128	40	900	12.553
SP NARINE	4146	691	200	4672	12.554

The data shows that the top 3 bowlers of IPL history are Lasith Malinga, Rashid Khan and Anil Kumble based on their CBR score. One thing to remember here is that the dataset contains bowlers who have retired from cricket. However, they are kept in the dataset to understand all bowler's performances in the IPL.

4. CONCLUSION

Advanced metrics like the Bayesian approach for batsmen and the harmonic mean (CBR) for bowlers offer valuable insights beyond traditional methods. The Bayesian approach tackles "not-out" innings, while CBR combines bowling metrics for a holistic view. This analysis revealed distinctions between top players, like Kohli edging Sharma in batting and Malinga, Rashid Khan and Kumble leading in bowling based on CBR.x

However, it is essential to recognize the limitations and contextual dependencies of these metrics. For batsmen, performance is influenced by various factors beyond runs scored, including player fitness, teamwork, and adaptability to different match situations. Similarly, while the harmonic mean approach provides a robust assessment of bowlers, it may not fully capture the specific demands of different bowling roles, such as death over specialists who prioritize economy rate. Other qualitative factors, such as the variety and effectiveness of deliveries, also play a crucial role in a bowler's impact on the game.

Future research should expand on these metrics by incorporating additional variables, such as player fitness, all-rounder capabilities, and psychological attributes, to provide a more comprehensive evaluation. The field of sports analytics, particularly in cricket, holds significant potential for growth. In India, where cricket talent is budding across the country, there is a substantial opportunity to leverage data analytics to unearth and develop talent, ultimately enhancing the sport's competitive landscape.

In conclusion, while the Bayesian and harmonic mean approaches offer innovative ways to measure player performance, they should be integrated with other metrics and qualitative assessments for a holistic understanding. This multidimensional approach will better serve the dynamic and multifaceted nature of cricket, fostering more strategic and informed decision-making in the sport.

5. REFERENCES

[1] Wood, G. H. (1945). Cricket scores and geometric progression. *Journal of Royal Statistical Society,108,* 12–22. Series A (Statistics in Society).

[2] Damodaran, U. (2006). Stochastic dominance and analysis of ODI batting performance: The Indian cricket team, 1989–2005. *Journal of Sports Science and Medicine, 5,* 503–508.

[3] Lemmer, H. H. (2002). The combined bowling rate as a measure of bowling performance in cricket. *South African Journal for Research in Sport, Physical Education and Recreation, 24*(2), 37–44.

[4] KENNEY, J.F. & KEEPING, E.S. (1954). *Mathematical Statistics* (3rd ed.). New York, NJ: D. van Nostrand.

[5] IPL matches dataset. https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020?select=matches.csv

[6] IPL ball-by-ball dataset. https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020?select=deliveries.csv.

[7] Fernández Ponce, J. M., Rodríguez Griñolo, M. del R., & Troncoso Molina, M. A. (2022). An Application of Stochastic Dominances in Sports Analytics. *Economía Aplicada, 40*(1). https://doi.org/10.25115/eea.v40i1.7002