International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 3 Issue 6‖ June 2014 ‖ PP.07-11

"Reflecting Against Perception: Data Analysis of IPL Batsman"

Amit Kumar[1], Dr. Ritu Sindhu[2,]
[1,2,]*Computer Science Engineering, SGTIET, Gurgaon/ Maharishi Dayanand University, Rohtak, India*

ABSTRACT: *This paper analyzes the (Performance of the batsman based upon the run scored different condition during IPL) using data of cricket, particularly IPL. It uses some predictive modeling to identify the trend and it explains different pattern developed by mining of that captured data. These pattern are either deviated or against our perception about them. This paper discuss various analysis created from IPL cricket data from 2008-2014. So in short it converts data into relevant information, which can help better understanding of the game & player's performance as raw data doesn't portrait the complete picture/aspect of any player.*

KEY WORD: *Data analytics, Data mining, sports data analysis, cricket, IPL, predictive modeling.*

I. INTRODUCTION

Cricket is played globally; It is the most popular game in India. It is played in various format like Test series[1,2], One day[3] and twenty- twenty (T-20)[4]. In twenty –twenty (T-20) is a recently developed format in which each team bat for twenty overs. It can be played in day and night format also. Due to its short duration it is becoming more popular than other format. Indian Premier League (IPL) which was started in year 2008[5], which is hosted (administrated by BCCI -Board of Cricket Control of India) is much popular than any other domestic tournament. Sportsmen from all over the world participate in this tournament, and viewers from all over the world follow the tournament.

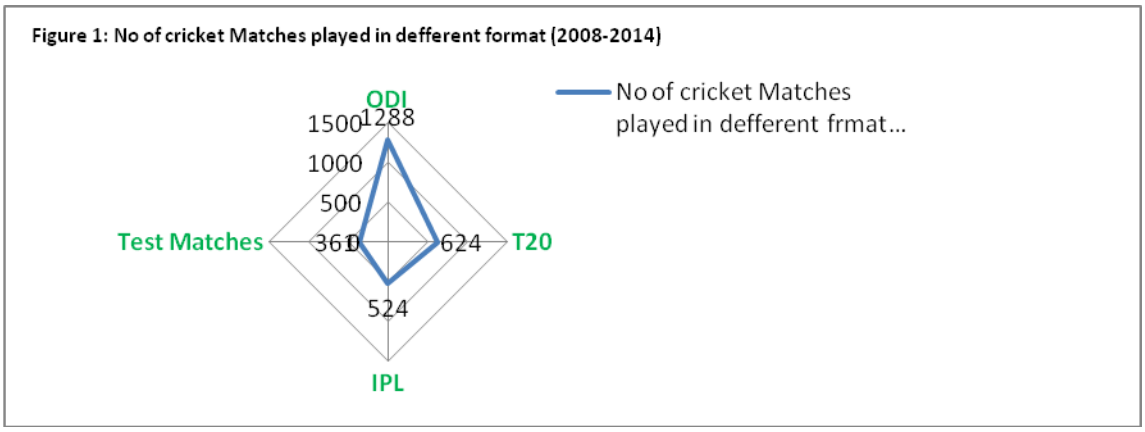

Figure-1: indicates that ODI and twenty –twenty format is played more than other format. IPL matches which is a approx 60 days tournament, in 5 years around 524 matches were played and number of run scored per match is higher than any other format.

Source: www.IPLt20.com, http://www.icc-cricket.com, http://www.espncricinfo.com/
This paper analyzes the data of IPL by creating different performance matrices it discuss some experiments that analyze batting performance of a given player. It also discusses the use of impact model matrices to analyze the performance of a batsman.

Table I: The IPL so far in numbers:

Year	*Matches*	*Runs*	*Wkts*	*RPO*
2014	60	18909	671	8.2
2013	76	22541	909	7.67
2012	75	22,453	857	7.82
2011	73	21,154	813	7.72
2010	60	18,864	720	8.12
2009	57	16,320	697	7.48
2008	58	17,937	689	8.3

Source: http://www.espncricinfo.com/, http://www.iplt20.com, Wkts:wickets, RPO:run rate per over

II. CAPTURING THE DATA

Secondary data has been taken from the official website of IPL T20[6], ICC [7] and espncricinfo[8]. A top twenty batsman data is captured and analyzed. Data has been collected from the year 2008(inception year) till 2014 (recently ended season) for table-3 data used is 2008-2012 taken from www.iplt20.com.

MATRIX

Unlike other conventional matrices that measure bating performance (such as Total no of runs, batting average, strike rate, etc) Continuous Adjusted Average and batting impact are used.

Continuous Adjusted Average

Continuous-adjusted average (CAA) Matrix is used to rank the top batsmen of all time in Test cricket. In this paper we have used it in the context of IPLs. As discussed in Borooah and Mangan[9], just looking at the average of a batsman over a period of time might not provide the full picture of how the batsman has performed[10]. Averages can be inflated because of a few high scores in non relevant matches or against weaker teams or non critical games. It is important to measure how consistent the batsman was in his performance. To measure Continuous, the Gini coefficient is used (Gini 1912)[11]. The Gini coefficient measures the inequality among values of a frequency distribution. The values can range from 0 to1. A low value indicates a more equal distribution (with 0 corresponding to complete equality), and a high value indicates a more unequal distribution (with1 corresponding to complete inequality). In the context of one-day international cricket, the Gini coefficient is computed as

$$G = 1\frac{1}{2N^2\mu}\sum_{i=1}^{N}\sum_{j=1}^{N}|Ri - Rj| \qquad (1)$$

Where *G* is the Gini coefficient, *N* is the number of complete innings, μ is the average score, and R is the number of runs scored by the batsman in the *i*th innings.

Then the Continuous-adjusted average (CAA) is computed as

$$CAA = \mu*(1-G) \qquad (2)$$

Where μ is the Continuous-unadjusted average (that is the regular average).

Batsman impacting Score (BIS)

Batsman impacting Score (BIS) evaluates performance of a player in reference to a match. It takes into account not only how many runs a player has scored but also the pace at which he scored the runs and the conditions under which he scored the runs. This is similar to some of the metrics discussed on the Impact Index website (Impact Index Cricket Pvt. Ltd.)[11]. A BI score is assigned to every player who batted in a given match based on the following aspects of his performance:

Runs Impact Score (RIS)

This metric measures the ratio of the runs scored by the player against the mean runs for all players for the match. The RIS for a player *p* in match *m* is computed as

$$RISpm = \frac{RUNSpm}{BASE\ RUNSm} \quad (2)$$

$$BASE\ RUNSm = \sum_{i=1}^{N} \frac{RUNSim}{N} \quad (3)$$

Where ***RISpm*** represents the runs scored by player *p* in match *m*, and *N* is the total number of players who batted in the match.

Strike Rate Impact Score (SRIS)

This metric assigns a positive score to the player if his strike rate is above the mean strike rate for the match and a negative score if it is below. If his strike rate equals the mean strike rate of the match, then his SRIS is equal to zero. The SRIS for a player *p* in match *m* is computed as

$$SRISpm = \frac{SRpm}{\frac{\sum_{i=1}^{N} RUNSim}{\sum_{j=1}^{N} BALLSim}} - 1 \quad (4)$$

Where ***SRISpm*** ***i***s the strike rate of player *p* in match *m*, and is the number of balls faced by player *I* in match *m*.

Pressure Impact Score (PrIS)

The score for this metric is a product of two variables:

(a) the situation in which the player comes in to bat and

(b) The player's RIS.

The first variable is called the Pressure Factor (PF), which depends on the difficulty of the situation when the batsman comes to bat. A situation's difficulty is measured in terms of the number of wickets that have fallen and the mean runs for the match. The PrIS for a player *p*in match *m* is computed as

$$PrISpm = PFpm * RISpm \quad (5)$$

$$PFpm = \frac{(INWpm * BASE\ RUNSm) - INSpm}{BASE\ RUNSm} \quad (6)$$

Where PFm is the number of wickets that have fallen in the innings when player P comes to bat in match m. is the score of team when player p comes to bat in match m.

If player p is an opening batsman, then it is always 0. To account for this all opening batsmen are given a PF of 0.5. In other words, he is expected to score at least half the number of mean runs for the match.

Chasing Impact Score (CHIS)

This special Score is assigned to a player for staying not out in the second innings of a successful chase. If a player satisfies this criterion then his CHIS is equal to his RIS, otherwise It is equal to 0.

Based on the preceding four scores, the overall batting impact score (BI) of a player p in a match is computed as

$$BIpm = \frac{RISpm + SRISpm + PrISpm + ChISpm}{MaxBIm} \quad (7)$$

Where BIpm is the highest batting score to match m and is used as a normalization factor.

Here it is obvious by table-2 that S K Raina whose is more consistent than Rohit G Sharma but by ranking of Average, he is no.2 but according to CAA it is oblivious that he is no 1. Similarly Utthappa, Marsh is over ranked, while SR Tendulkar, G Gambhir is under ranked due to experimental error in Average, although they have consistently performed during the season.

Table -II: The Analysis of Ranking of players who played in the year 2008-14

S. No.	*Name of player*	*Rank by Avg.*	*Average*	*Continuous adjacent average(CAA)*	*CAA*	*Difference in rank.*
1	S K Raina	2	51.84	39.98	1	1
2	R G Sharma	1	52.95	38.09	2	-1
3	G Gambhir	6	47.00	32.34	3	3
4	C H Gayle	3	51.01	31.98	4	-1
5	V Kohli	7	46.30	30.58	5	2
6	P V utthapa	4	47.68	28.54	6	-2
7	V Sehwag	8	45.54	27.89	7	1
8	MS DHONI	10	42.10	26.00	8	2
9	JH KALLIS	5	47.45	24.70	9	-4
10	SR TENDULKAR*	12	42.08	24.20	10	2
11	S DHAWAN	9	44.02	23.73	11	-2
12	RDRAVID*	11	42.44	23.03	12	-1
13	Y. PATHAN	18	34.89	22.06	13	5
14	AC GILCHRIST *	17	35.09	21.45	14	3
15	KD KARTHEE	13	41.21	20.68	15	-2
16	AB D-VILLIARSA	14	39.92	20.66	16	-2
17	SR WATSON	16	35.07	20.33	17	-1
18	MARSH	15	38.74	19.57	18	-3
19	DA WARNER	20	33.59	19.46	19	3
20	HUSSEY	19	34.08	19.26	20	-1

Table III: Batting Ranking Matrices of players

S.No.	**Name of player***	**Average**	**Rank by Avg.**	**Continuous adjacent average BI**	**Rank by CABI**	**Difference in rank.**
1	Chris Gayle	52.28	**2**	65.54	**1**	1
2	Virender Sehwag	53.85	1	46.70	2	-1
3	Shane Watson	47.00	6	43.23	3	3
4	David Warner	51.65	3	39.42	4	-1
5	Shaun Marsh	46.89	7	36.03	5	2
6	Gautam Gambhir	48.67	4	33.85	6	-2
7	Shikhar Dhawan	45.02	8	33.30	7	1
8	Kevin Pietersen	43.80	9	31.25	8	1
9	Sachin Tendulkar	47.55	5	31.09	9	-4
10	Suresh Raina	42.37	10	29.84	10	0

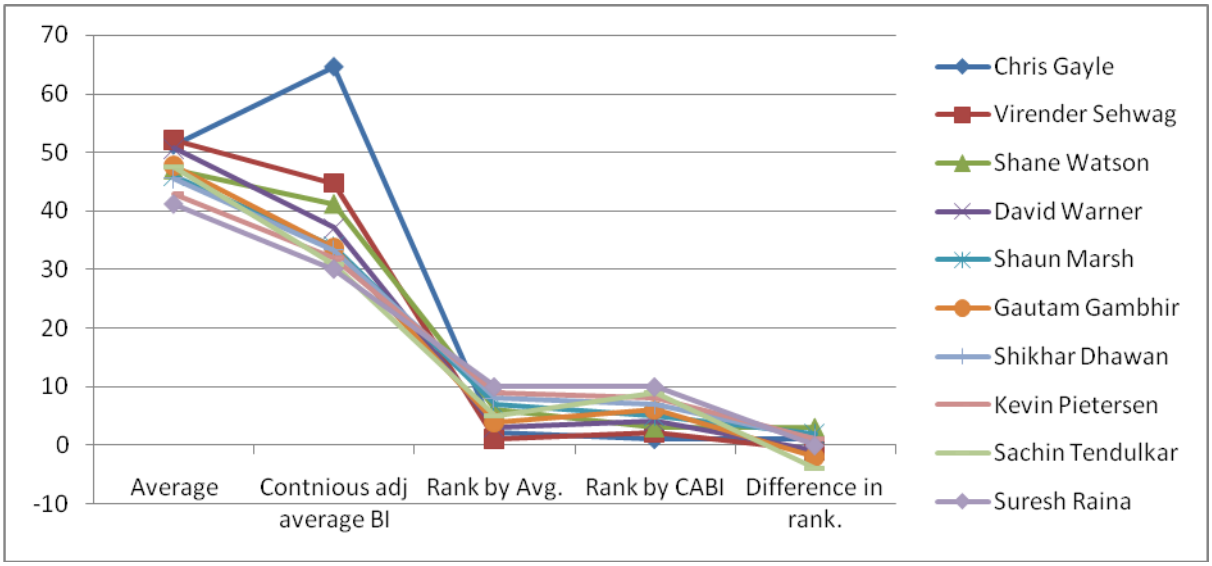

Fig.2 Graph for Batting Ranking of Players from Table-III

Graphical representation of data in any sports[12] From Table 3 it is obvious that Chris Gayle who has been ranked two by average system have performed consistently and he has been a match winner. He has performed in critical matches second surprise is SR Tendulkar who failed to click in crucial matches. Same thing is with Gautam Gambhir. He has performed below par in critical matches from 2010-2012 in IPL matches

III. CONCLUSION

From the above discussion it can be concluded that there are various factor other than average which contribute to the find out real performance indicator. So while analyzing any data we should use advance matrices to get the real projection. The consistent performance is more important than any other random score. but there are limitation for any predictive and performance model, so all possible factor should be factor in before reaching on any conclusion.

REFERENCES

[1] Classification of Official Cricket
[2] Standard test match playing conditions http://icclive.s3.amazonaws.com/cms/media/about_docs/524ac43be2acaStandard%20Test%20Match.pdf
[3] Standard ONE-day International match playing conditions http://icc-live.s3.amazonaws.com/cms/media/about_docs/524ac4ae08b48-04%20Standard%20ODI_2013_19%2009%2013.pdf
[4] Standard TWENTY20 International match playing conditions http://icc-live.s3.amazonaws.com/cms/media/about_docs/524ac4f092e53-05%20Standard%20T20_2013_19%2009%2013.pdf
[5] History: all about IPL http://sports.in.msn.com/cricket/ipl-5/history/aboutipl.aspx
[6] http://www.iplt20.com/
[7] http://www.icc-cricket.com/
[8] http://www.espncricinfo.com/
[9] Borooah, V.K. and Mangan, J., (2012). "Measuring Competitive Balance in Sports using Generalized Entropy with an Application to English Premier League Football," Applied Economics, 44, 1093-1102.
[10] Analysis of Batting Performance in Cricket using Individual and Moving Range (MR) Control Charts: Muhammad Daniyal1,Tahir Nawaz1, Iqra mubeen 2, Muhammad Aleem, ISSN 1750-9823 (print)International Journal of Sports Science and Engineering, Vol. 06 (2012) No. 04, pp. 195-202.
[11] Inequality Analysis, The Gini Index: Lorenzo Giovanni Bellù, Agricultural Policy Support Service, Policy Assistance Division, FAO, Rome, Italy https://www.princeton.edu/~achaney/tmve/wiki100k/docs/Gini_coefficient.html
[12] Comparison of bowlers, batsmen and all-rounders in cricket using graphical displays, Paul J. van Staden technical report, University of Pretoria, Faculty of natural and agricultural sciences, Department of statistics, ISBN: 978-1-86854-733-3