International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-9, Issue-9, September 2023
ISSN: 2395-3470
www.ijseas.com

Cricketing Performance Analysis in IPL 2023: Unravelling Key Influencing Factors through Principal Component Analysis and Biplot Clustering

D S Dhakre and Debasis Bhattacharya
Department of Agricultural Statistics, PSB, Visva-Bharati, Sriniketan, WB
Corresponding Author email: dhakreds@gmail.com

ABSTRACT

This study is performed to identify the cricketing performance measures which is significantly influence players' performance in IPL 2023 using Principal Component Analysis. This article aims to simplify the methods of finding the cluster of players for in PCA biplot in this study the data of players who performed in IPL 2023 are clustered based on cricketing performance measure by using two principal axes PCA biplot it is found for IPL 2023 data that the players are clustered into two groups. Out of these two groups first group shows the player with high performance, especially Run, number of 4s, number of 6s, numbers of 50s. and the second group shows the number of wickets for the bowlers.

KEY WORDS:
Cricket, Cricketing measures, Multivariate analysis (MVA), Principal component analysis

INTRODUCTION

The role of multivariate analysis (MVA) in data mining is very prominent and most of the multivariate analysis techniques are designed to summarize information which remain unknown or hidden in a large data set. Data mining is the art of extracting information using different statistical techniques in general and MVA techniques as particular. Multivariate analysis techniques are often used in exploratory data analysis (EDA) or graphical data analysis (GDA). Multivariate analysis used in EDA includes factor analysis, cluster analysis, discriminant analysis, correspondence analysis (CA), Principal Component Analysis (PCA), logistic regression, etc.

Principal component analysis is used to reduce the dimensionality of a multivariate data set. In this technique correlations and interactions among the variables are summarized in terms of a small number of under playing factors. The method rapidly identifies the key variables or groups of variables that control the system under study. Principal component analysis of a set

of 'm' original variables generate 'm' principal components, PC_1 , PC_2 , ..., PC_m , where each principal component being a linear combination of PCs' scores on the original variable,

$$PC_1 = b_{11} X_1 + b_{12} X_2 + \ldots + b_{1m} X_m = X_{b1} ;$$

$$PC_2 = b_{21} X_1 + b_{22} X_2 + \ldots + b_{2m} X_m = X_{b2} ;$$

$$PC_m = b_{m1} X_1 + b_{1m} X_2 + \ldots + b_{mm} X_m = X_{bm} ;$$

where, b_i refer to the coefficients, i = 1 to m

X_i refers to the i^{th} variables, i = 1 to m

The coefficients for PC_1 are chosen to make its variance as large as possible. The coefficients for PC_2 are chosen to make the variance of this combined variable as large as possible, subject to the restriction that scores on PC_1 and PC_2 (whose variance has already been maximized) are uncorrelated. In general, the coefficients for PC_1 are chosen to make its variance as large as possible, subject to the restriction that it will be uncorrelated with scores on PC_1 through PC_{i-1}.

MATRIALS AND METHODS

The real data set has been taken from https://www.ipl20.com. A data set of batsman's consists of 166 batters with 7 bating measures sports parameters such as Run, Batting average (Avg), Strike Rate (SR), number of hundred's (X100s), number of fifty's (X50s), Number of four (X4s), number of six of baters (X6s) while data set of bowlers consists of 113 bowlers with 4 bowling measures such as Wickets (Wkts), Bowling average (Avg), Economic bowler (Econ), Strike Rate (SR). PCA requires quantitative scale variables, so the data set prepared appropriately before analyzing the data.

RESULTS AND DISCUSSION

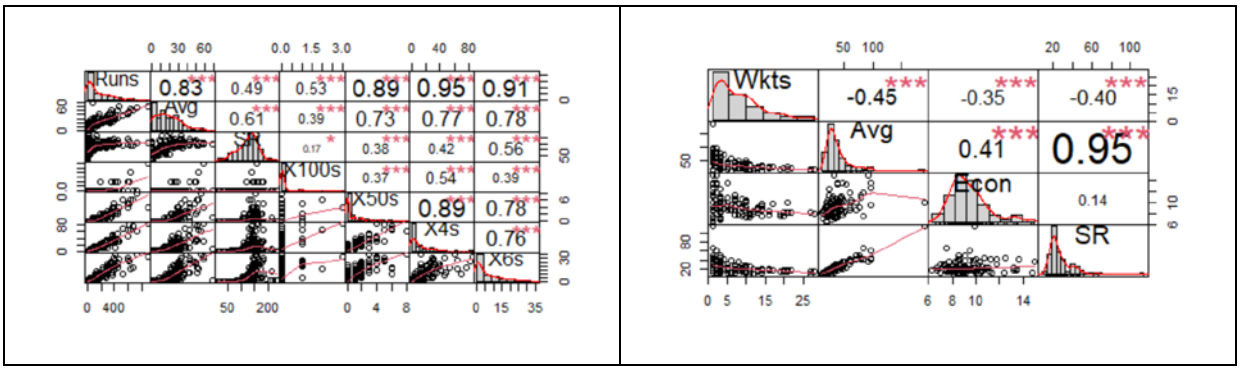

Fig1: Correlation Coefficient of measures of cricket for Batters and Bowlers

Several sports measures are found to be highly correlated (Fig.1), and the fig shows that multicollinearity is present in the data. the important tests for PCA the Kaiser-Meyer-Olkin (KMO) of sampling adequacy and Bartlett's test of sphericity. The following table (Table 1) gives the following values: KMO is 0.762 for Batters and 0.367 for Bowlers. Barlett's test has been found to be highly significant ($p < 0.001$) for both and therefore factor analysis is considered appropriate for analyzing this data

Table 1: KMO and Bartlett's Test

		Baters	Bowlers
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.762	0.367
Bartlett's Test of Sphericity	Approx. Chi-Square	1444.771	478.108
	df	21	6
	Sig.	0.00	0.00

The number of principal components to be considered is determined by looking at the eigenvalues which are greater or equals 1. From Table 2, the percentage model quality for 1st PC is respectively 70.09% for Batters and 60.45% for Bowlers.

Table 2. Eigen values and proportion of the variance of principal components (PC)

	Batsmen		Bowlers	
Component	PC1	PC2	PC1	PC2
Eigenvalues	4.91	0.89	2.42	0.95
% of Variance	70.09	12.72	60.45	23.75
Cumulative %	70.09	82.81	60.45	84.20

Table 3. Orthonormal eigenvector matrix (component coefficient).

	Runs	Avg	SR	X100s	X50s	X4s	X6s
Batsmen	0.443	0.401	0.272	0.249	0.403	0.422	0.408
	Wkts	Avg	Econ	SR			
Bowlers	0.435	-0.613	-0.34	-0.565			

The principal component based models for Batters and Bowlers are given as follows:

$Z_{batters}$ = 0.443*Run+0.401*Avg +0.272*SR+0.249*X100s+0.403*X50s+0.422*X4s + 0.408*X6s

and $Z_{bowlers}$ = 0.435*Wkts -0.613*Avg - 0.34*Econ - 0.565*SR

Fig.2 gives the scree plot of the principal components which helps in determining graphically the number of PCs to be considered significant and sufficient for the underlying data

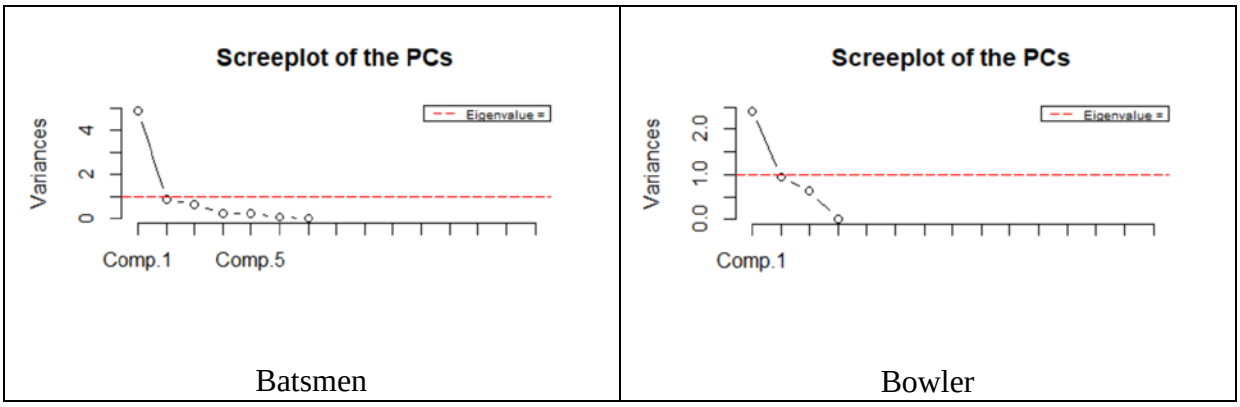

Fig 2: Scree plot of PC's for Batters and Bowlers

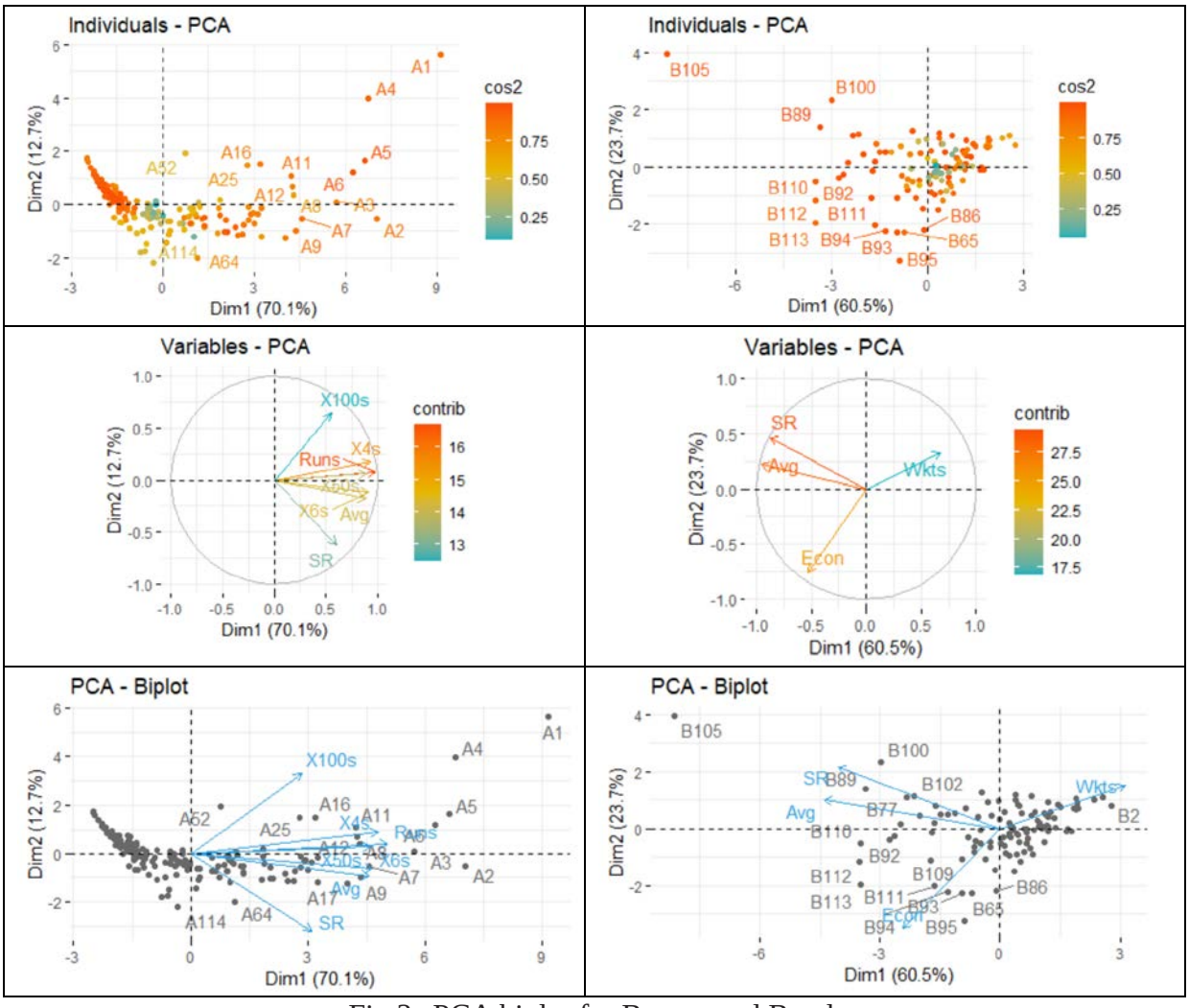

Fig.3. PCA biplot for Batters and Bowlers

Fig.3 presents the data visually. The obtained information's of players can be classified into four groups based on the quadrant. The first group is the players in quadrant (upper right), that are the players with high performance in Bating based on Run, X4s, X6s, X50s. and Bowlers based on Wkts.

Table 4: Top performed Batsmen in IPL 2023 as per First Principal Component Z_{Batter}

Player Code	Player	Runs	Avg	SR	X100s	X50s	X4s	X6s	Z_{Batter}
A1	Shubman Gill	890	59.33	157.8	3	4	85	33	9.10
A2	Faf Du Plessis	730	56.15	153.68	0	8	60	36	7.02
A4	Virat Kohli	639	53.25	139.82	2	6	65	16	6.74
A5	Yashasvi Jaiswal	625	48.08	163.61	1	5	82	26	6.59
A6	Suryakumar Yadav	605	43.21	181.13	1	5	65	28	6.21
A3	Devon Conway	672	51.69	139.7	0	6	77	18	5.70
A7	Ruturaj Gaikwad	590	42.14	147.5	0	4	46	30	4.56
A9	Rinku Singh	474	59.25	149.52	0	4	31	29	4.34
A8	David Warner	516	36.86	131.63	0	6	69	10	4.30
A12	Heinrich Klaasen	448	49.78	177.07	1	2	32	25	4.24

Table 5: Top performed Bowlers in IPL 2023 as per First Principal Component Z_{Bowler}

Player Code	Player	Wkts	Avg	Econ	SR	Z_{Bowler}
B2	Mohit Sharma	27	13.37	8.17	9.81	2.74
B1	Mohammad Shami	28	18.64	8.03	13.92	2.53
B3	Rashid Khan	27	20.44	8.23	14.88	2.34
B5	Yuzvendra Chahal	21	20.57	8.17	15.09	1.93
B9	Mohammed Siraj	19	19.78	7.52	15.78	1.92
B4	Piyush Chawla	22	22.5	8.11	16.63	1.90
B7	Ravindra Jadeja	20	21.55	7.56	17.1	1.88
B10	Matheesha Pathirana	19	19.52	8	14.63	1.88
B8	Varun Chakaravarthy	20	21.45	8.14	15.8	1.82
B29	Mark Wood	11	11.81	8.12	8.72	1.75

CONCLUSION

The results of the principal component analysis of cricketing measure revealed that several measures are highly correlated as it shows the presence of multicollinearity in the data set.

Eigen values (>1) and the scree plot methods are used to identify the principal components to be retained as predictors. According to principal component analysis first component explained variation 70.09% for Batters and 60.45% for Bowlers of the original variables.

As per Biplot, players Shubman Gill, Faf Du Plessis, Devon Conway, Virat Kohli, Yashasvi Jaiswal, Suryakumar Yadav, Ruturaj Gaikwad, David Warner, Rinku Singh with high performance in Bating based on Runs, X4s, X6s, X50s. and players Mohammad Shami, Mohit Sharma, Rashid Khan, Piyush Chawla, Yuzvendra Chahal, Tushar Deshpande,

Ravindra Jadeja, Varun Chakaravarthy, Mohammed Siraj, Arshdeep Singh with high performance in Bowling based on Wickets, Avg, Econ, SR.

Reference

1. Ananda, B. W. Manage and Stephen, M. Scariano (2013). "An Introductory Application of Principal Components to Cricket Data", Journal of Statistics Education, 21
2. https://www.iplt20.com/stats/2023