Jaithra Bhatia
u1193694

Final Report: Cricket Data analysis on Batsmen

Abstract

Sophisticated data analysis approaches, such as data mining, where the emphasis is on exploration and discovering new insights, are becoming more effective tools in analyzing top sports performance data and aiding critical decision making. In this report, we look at the various data mining demands in the sports of Cricket in relation to a variety of variables that define the performances of Batsmen. The goal is to integrate the Cricket and data mining domains more structurally by (a) establishing a framework for categorizing different metrics and (b) comprehending the analytical methods used to create a ranking system for the batsmen. As a result, we examine the factors of performance analysis needs that impact each stage of Cricket data mining.

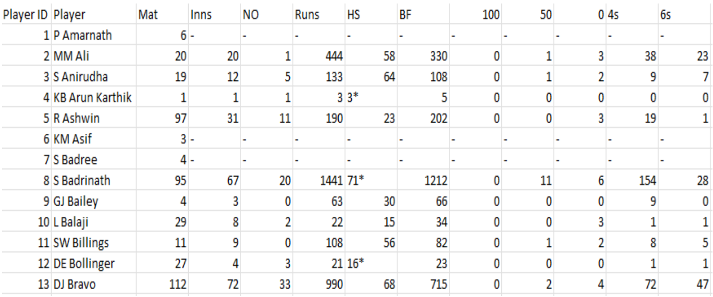
Introduction

Cricket is a game of numbers, such as runs scored by a batter, wickets taken by a bowler, matches won by a cricket team, and the number of times a batter responds to certain bowling attacks in a particular way. Cricket analytics could help provide interesting insights into the game and predictive information about the outcome of the game.

The Indian Premier League is a professional men's Twenty20 cricket league, contested by ten teams based in ten Indian cities. The league was founded by the Board of Control for Cricket in India in 2007. In the IPL Auction of 2022, 204 players were sold and INR 5,51,70,00,000 was splurged amongst the ten franchises during the two-day auction in Bengaluru. As IPL continues to grow as a League watched around the world, teams need to look at making purchases on players who will be able to contribute to their success. Performing data analysis on the past performance of players would help teams make the best choices. This report will be used to gather these player data and use clustering techniques to help rank the batsmen according to their performances,

Data collection process

Data for the batsmen's performances were gathered through ESPN Cricinfo. They have gathered data on all aspects of batsmen, such as the runs they have scored, balls they have faced, innings played, etc. This data was then exported into a CSV file through which I was able to create a table for the data.



Player ID	Player	Mat	Inns	NO	Runs	HS	BF	100	50	0	4s	6s
1	P Amarnath	6	-	-	-	-	-	-	-	-	-	-
2	MM Ali	20	20	1	444	58	330	0	1	3	38	23
3	S Anirudha	19	12	5	133	64	108	0	1	2	9	7
4	KB Arun Karthik	1	1	1	3	3*	5	0	0	0	0	0
5	R Ashwin	97	31	11	190	23	202	0	0	3	19	1
6	KM Asif	3	-	-	-	-	-	-	-	-	-	-
7	S Badree	4	-	-	-	-	-	-	-	-	-	-
8	S Badrinath	95	67	20	1441	71*	1212	0	11	6	154	28
9	GJ Bailey	4	3	0	63	30	66	0	0	0	9	0
10	L Balaji	29	8	2	22	15	34	0	0	3	1	1
11	SW Billings	11	9	0	108	56	82	0	1	2	8	5
12	DE Bollinger	27	4	3	21	16*	23	0	0	0	1	1
13	DJ Bravo	112	72	33	990	68	715	0	2	4	72	47

Methods and materials

In IPL, you need batsmen who are capable of producing a high rate of runs and good consistency of runs in these limited-overs provided. Using the data provided, I then had to create a metric system to help formulate a way to cluster them. These are the metrics that I used:

Hitting Ability, how many 4s and 6s can batsmen hit in the number of balls faced,

Strike Rate, the ability to quickly convert runs in the balls faced, and,

Consistency in the total runs they have scored in the number of innings they have played.

To help me create a dataset containing only the high-performing batsmen according to these metrics, I then went ahead and created a distribution graph using the Histogram Method to help me get rid of a few outliers and set a lower bound for these metrics.

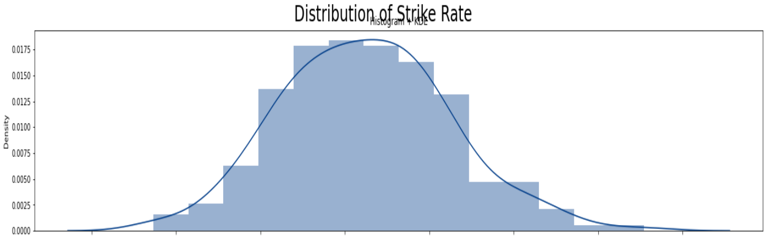


Figure 1: Histogram before Outliers

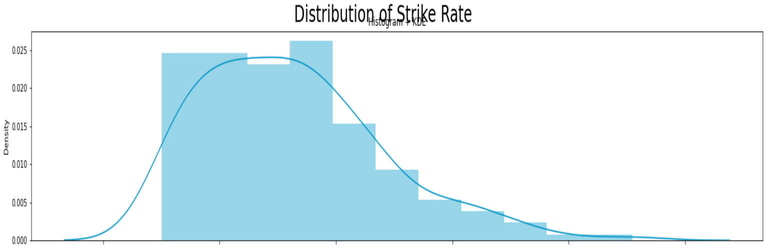


Figure 2: Histogram after Outliers

Then I perform the elbow method for all 3 metrics to give me the ideal number of clusters.

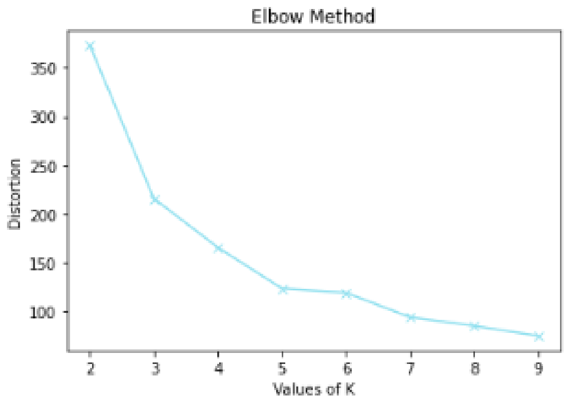


Figure 3: Elbow on "Average", "Strike_Rate"

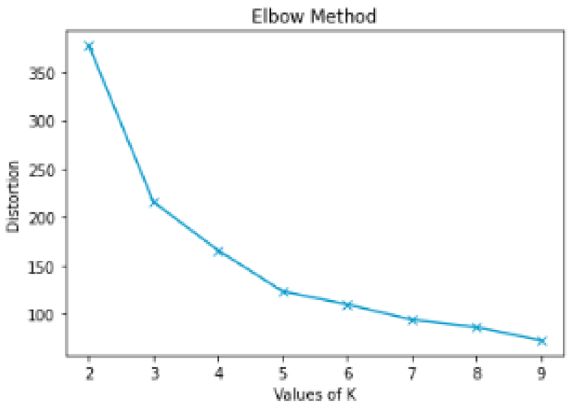


Figure 4: Elbow on "Average", "Hitting Ability"

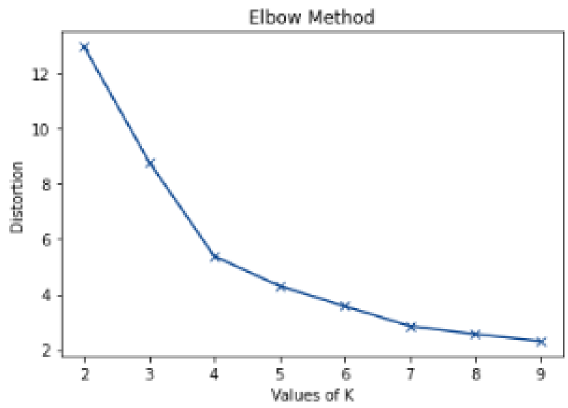


Figure 5: Elbow on "Strike_Rate", "Hitting Ability"

Results

We first present the clustering done for the metrics in a 2-D graph.

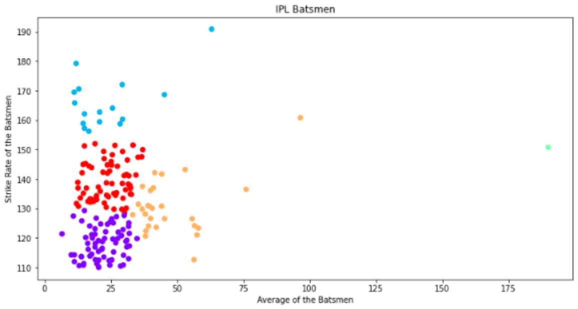


Figure 6: Clustering on "Strike_Rate", "Average"

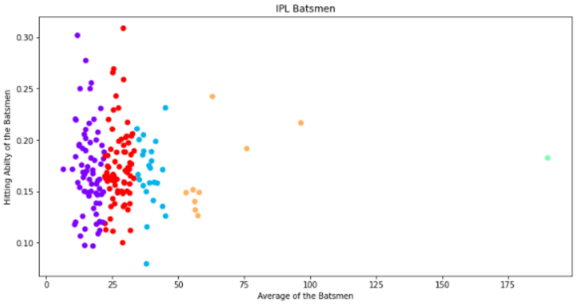


Figure 7: Clustering on "Hitting Ability", "Average”

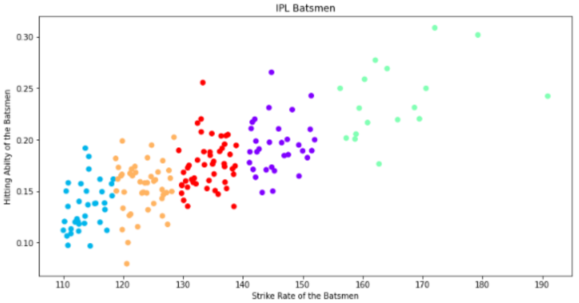


Figure 8: Clustering on "Hitting Ability", "Strike_Rate”

From the Clusters, we can see that even though the values in Figures 6 & 7 seem to be similar, the clusters formed on them are vastly different. Figure 8 has a nice spread of clusters which helps us better rank players in the aspect of Hitting Ability and Strike Rate. This helps us understand that players who have a higher Strike Rate also tend to have a higher hitting ability. Figure 8 produces a graph that would be more beneficial to the team drafting high-rated batsmen.

I also create a 3-D graph using all 3 metrics for each axis.

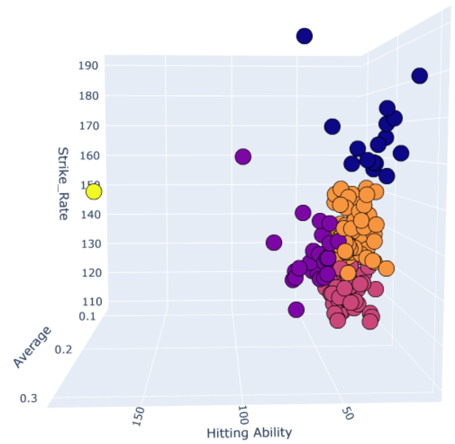


Figure 9: 3-D Clustering

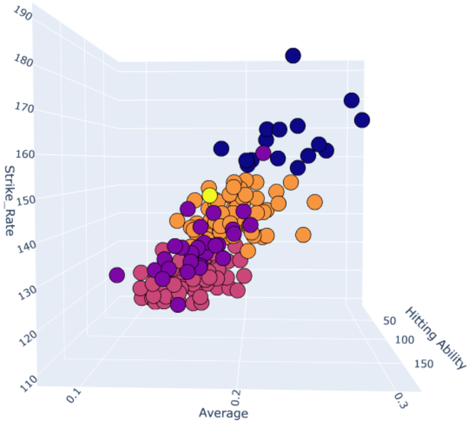


Figure 10: 3-D Clustering

Figures 9 and 10 helps creates clusters ranking the players according to all 3 metrics. According to these figures, we can see the blue cluster which shows us batsmen who rank greatly in all 3 metrics. This cluster consists of players who have also been sold for big amounts in previous IPL seasons.

Conclusion

Using the information gathered in this paper, we can distinguish batsmen in different rankings according to the metrics created, and then use those rankings to help teams understand how much budget can they then use to draft batsmen which will help teams perform better.

Citations:

"Get the Latest Auction Details of Every IPL Team 2022." *Get the Latest Auction Details of Every IPL Team 2022| IPLT20.Com*, https://www.iplt20.com/auction.

Naren. "Kmeans Clustering and Cluster Visualization in 3D." *Kaggle*, Kaggle, 11 Sept. 2019, https://www.kaggle.com/code/naren3256/kmeans-clustering-and-cluster-visualization-in-3d/notebook.

"Data Mining in Elite Sports: A Review and a Framework." *Taylor & Francis*, https://www.tandfonline.com/doi/full/10.1080/1091367X.2013.805137.