ClassicXI – Optimal IPL Team Selection using Machine Learning Models

Lavitra Kshitij Madan
CSE Dept
PES University
Bangalore, India
lavitra.kshitij@gmail.com

Ritik Hariani
CSE Dept
PES University
Bangalore, India
ritikhariani@gmail.com

Arjun Chengappa
CSE Dept
PES University
Bangalore, India
arjunchang25@gmail.com

Aditya G Burli
CSE Dept
PES University
Bangalore, India
adityaburli06@gmail.com

***Abstract*—Indian Premier League (IPL) is one of the most prestigious T20 cricket tournaments across the world. This paper focuses on rating different players using an ML model and then given a playing squad, provide an optimal playing 11 team out of it, based on strength (Yet to be updated based on the machine learning model that needs to be chosen such as Regression Testing feasibility, KNN Implementation feasibility, Decision Tree exploration feasibility, Random Forest feasibility and Naïve Bayes feasibility)**

Keywords—Cricket, IPL, Exploratory Data Analysis, Machine Learning

I. INTRODUCTION

The Indian Premier League (IPL) is an annual T20 tournament held in India. It was started in 2008 by the BCCI and it is now in its 13th season. It is now considered to be the best T20 tournament in the world. The success of the IPL can be attributed to a number of things, but one substantial factor to it was the quality of cricket it produced, hence providing quality entertainment for the audience. The IPL teams are owned by franchises that participate in auctions to purchase players- both domestic and international. The auctions are held every year, where players released from the teams and other uncapped players are put up to bid. A team is allowed to have a maximum of 25 players in its squad. The playing eleven of a team must have a maximum of four overseas players. The auctions are held every year, where players released from the teams and other uncapped players are put up to bid.

Every team spends a large amount of time and money to decide which eleven players to select from the 25 in the bag. The T20 format is a peculiar one, where the game is slightly skewed in favour of the batsmen compared to the bowlers. Shorter boundaries, a 6 over power play, day/night conditions etc. make up the reasons for that. Hence, there is no doubt that the teams give emphasis on the strength of their batting line up. Teams tend to have a mix of batsmen that can score at a good strike rate and those that bring stability to prevent a collapse. The order in which the two kinds of batsmen are sent also makes a huge difference. The opening pair generally has one batsman that can be explosive and another that takes a few balls to get going. The middle order is considered to be very crucial in T20 and it generally has a combination of experienced batsmen that can take the game till the end and batsman who can score big from the first ball. All-rounders are also known as game changers in the T20 format, as they can contribute with either their batting or bowling. Teams prefer to have a mix of batting and bowling all-rounders, and it is the all-rounders of a team that are generally juggled around in accordance to the match being played.

Since the game is skewed in favour of batting, it also becomes very important for a team to have a potent and balanced bowling line up. In the modern game, the diversity of bowlers is far greater. The line-up needs to have a mix of spin and seam bowling; impact bowlers that can either take wickets with the new ball, take control of the middle overs or restrict the opposition in the death overs. In the IPL, teams also have to decide on which four overseas players get their place in the playing eleven.

Given all this, we tend to notice that the process of selecting a playing eleven requires the teams to analyse all of their players individually to figure out whether they make the cut and where they are to be placed. Data Analytics and mathematical modelling can be extensively used in the process. This paper performs various exploratory data analysis (EDA) on a couple of datasets- the ball by ball dataset of the IPL and the match dataset to derive at required intuitions and deeper understanding of the game that will facilitate in the final effort to build a comprehensive model to compute the playing eleven of an IPL team given a particular squad. EDA is performed on different aspects of the game such as- analysis of fielding and dismissals, toss analysis, average of batsmen against a team, average wickets of a bowler against a team etc. The data gathered from this will be used to compute ratings of all the players, which in turn will help in deciding the playing eleven of a squad.

II. RELATED WORK

A lot of statistical analysis has been done over the last decade on cricket. There has been an increase in the attention of performing analysis on the Indian Premier League due to the large sum of money invested in it. This had led to researchers taking the initiative of predict match outcomes, predicting auction sale prices of various cricketers. The respective IPL teams' management also include of data analysts to provide input to the team's owners and coach on strategies on which player to choose and for what price. They also play a significant role in suggesting improvements in the various areas where performance is weak such as fielding, a batsman's strike rate, a bowler's economy, player's fitness levels based on the consumption of food and his physical training activities.

[1] have devised a method that ranks players based on the following factor - Deep Performance Index method to evaluate a player. The algorithm considers context of the player in terms of indices like most valuable player index for batsman and for bowler. The important context introduced is the winning contribution ratio of the player. An approach for evaluating a player based on the venue of match, overall average, current season average and additional parameters. Batting Rating Points System (**BRPS**) = **WCR** + **AVI** + (**BCSA** / **BOA** * **BCSSR** / **BOSR**) + **HPR**,

where, WCR is the Winning Contribution Ratio, AVI is the Average Venue Index, BCSA is the Batsman Current Season Average, BOA is the Batsman Overall Average, BCSSR is the Batsman Current Season Strike Rate, BOSR is the Batsman Overall Strike Rate, HPR is the Hitting Power Ratio

Bowling Rating Points System (**BoRPS**) = **WCR** + **AVI** + (**BoOA** / **BoCSA** * **BoOE** / **BoCE**) + **WMR**,
where, WCR is the Winning Contribution Ratio, AVI is the Average Venue Index, BoCSA is the Bowler Current Season Average, BoOA is the Bowler Overall Average,
BoOE is the Bowler Overall Economy, BoCE is the Bowler Current Season Economy, WMR is the Wickets Matches Ratio. The values of BPRS vary from 3.47 to 8.42 and that of BoRPS vary from 2.58 to 6.54.

Although, to apply the above algorithms the following assumptions were necessary: Factors involving team support and other factors such as fielding support or assists were ignored (Team Contribution). [1] Only considered IPL performance data up to IPL 7 and the overall T20 career data up to the end of IPL 8 of all the players participating in IPL. Batsman scoring more than 500 runs in twenty-20 internationals with a strike rate exceeding 100 and who have played in at least 25 matches were considered for batting ranking. Applying these criteria, 89 batsmen were considered for this rating system.

[2] provides insights about modelling batsmen, bowlers and ultimately modelling teams (rating) and suggesting a 'Playing XI', given a squad/pool of players using a few features. [2] propose novel methods to model batsmen, bowlers and teams, using various career statistics and recent performances of the players. They also propose that their model can predict the winner of ODI cricket matches, by using a novel dynamic approach to reflect changes in player combinations. A lot of assumptions have been made in the modelling step of batsmen, bowlers and the teams as a whole. [2] use features such as **Matches Played** (Matches played by the player), **Batting Innings** (Matches in which the player batted), **Batting Average** (Runs scored divided by the times the player got out), **Num Centuries** (Times the player scored ≥ 100 runs in a match), **Num Fifties** (Times the player scored ≥ 50 but less than 100 runs in a match), **Bowling Innings** (Matches in which the player bowled), **Wkts Taken** (Wickets taken by the player), **FWkts Hauls** (Times the player has taken ≥ 5 wickets in a match), **Bowling Average** (Runs conceded by the player per wicket taken) and **Bowling Economy** (Economy Average runs conceded by the player per over). Many important features such as venue conditions and relative strength within the team have been ignored. Another important concept is that all-rounders have not been considered separately but they are taken as batsmen or bowlers. Hence, our paper aims to also model all-rounders to improve the playing XI prediction quality. Our model provides more value to recent performances as the players learn and grow. Although [2] has been implemented for ODI cricket matches, IPL T20 would be a more challenging task and hence would be more valuable in terms of need in today's world. [2] claim that the k-Nearest Neighbor (kNN) algorithm used provides better results with an accuracy of 71% as compared to other classifiers models that have an accuracy of 56% and 63%.

[3] provides insight on how to tackle modelling all-rounders which was not taken care of in [2]. The paper takes into consideration that the player was in the playing eleven for at least 5 matches in the IPL, the player has bowled for at least 10 overs in IPL, the player has faced at least 100 balls in the IPL, and, only a total of 35 players have been considered for the research. Therefore, it does not represent the whole population of players. From this paper we learn that use of Wald statistic, Step-wise multinomial logistic regression (SMLR) and the naïve Bayesian classification model for forecasting the expected class of all-rounders based on the significant predictors can be a method to approach our problem. [3] based on the above method have claimed that all-rounders have been successfully classified into 4 non-overlapping categories, namely- Performing all-rounders, batting all-rounders, bowling all-rounders and Under-performer with an accuracy of their model at 66.7%.

III. METHODOLOGY

A. Dataset

The datasets used in this project are taken from Kaggle and is available at the following link: https://www.kaggle.com/nowke9/ipldata

This consists of two .csv files deliveries.csv and matches.csv. The file deliveries.csv contains the ball by ball data of every match since 2008. Data includes columns such as bowling team, batting team, over, ball, batsman etc. This dataset has a total of 21 columns and 179078 rows. The file matches.csv contains details about the match as a whole. Data includes columns such as season, city, teams, toss_winner, toss_decision, etc.

B. Pre-processing

The file deliveries.csv has at least 170244 NaN values in the attribute's fields: player_dismissed, dismissal_kind and fielder after cleaning.

The file matches.csv was fact checked and cleaned by the following steps,

All the empty places were replaced with NaN to avoid discrepancy and have an overall sense of consistency.

Some stadium names in this dataset had minute spelling mismatches. These were found when the unique values of the stadium attribute were sorted. For example, we found two names that pointed to the same stadium **'M. Chinnaswamy Stadium'** and **'M Chinnaswamy Stadium'**. This was resolved by individually replacing all the errors using the replace function of the 'pandas' library.

In some of the records it was observed that the number of innings was reported to be 5. This is not possible in T20 and is obviously an error in data entry. There were also cases where a match was reported to have 4 innings but weren't reported as super-overs. These errors were individually looked into and rectified.

A few discrepancies regarding no balls and wide runs format were observed due to the ever-changing laws of cricket. These were simply handled by a series of well-informed if-else statements.

IV. DATA ANALYSIS

To understand the data, and the features that will be needed to play an important role in our model, we have performed Exploratory Data Analysis and Visualizations to select out the important features.

The following graph Figure 1 shows the head to head data analysis for the team Royal Challengers Bangalore (number of wins, losses and number of tied matches as respective percentages against every other team they have faced in the Indian Premier League)

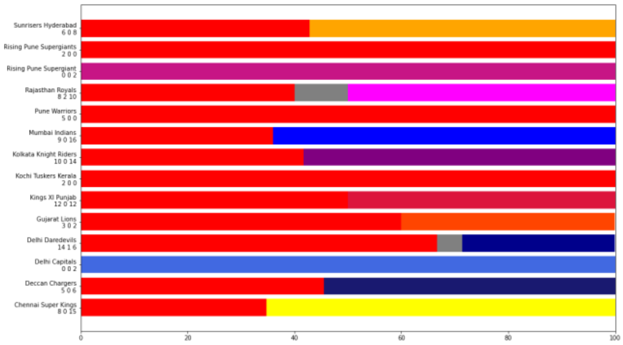


Figure 1

Figure 2. shows the toss analysis for the various venues on which the matches were played. It shows the win prediction possibility based on the data over 11 seasons of the IPL. It shows the probability that a team would likely win at a particular venue based on the toss decision (either choosing to bat or bowl first).

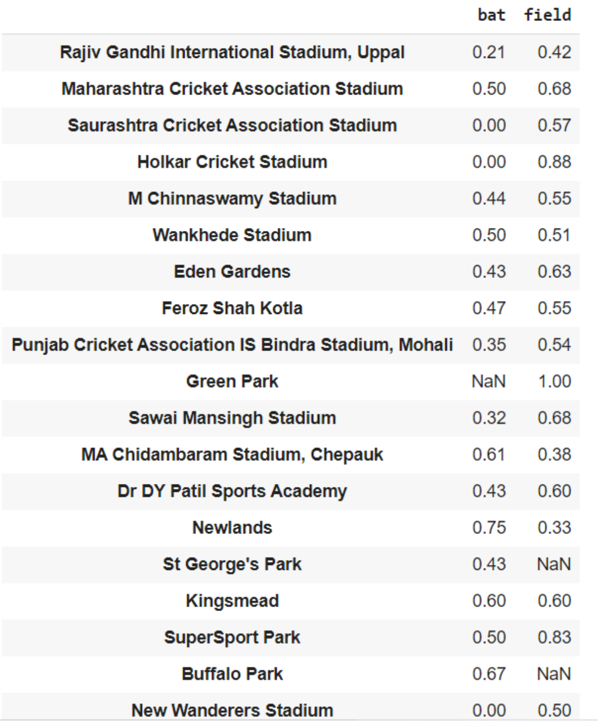
Figure 3. shows the number of dismissals categorized into 3 types – Caught, Stumped, Run out

Based on this we perform analysis on the number of times a batsman has got 'out' on each type of dismissal to understand his weakness and the area on which he must work up on.

Figure 4. shows the most caught for dismissal analysis of 10 players in the history of IPL

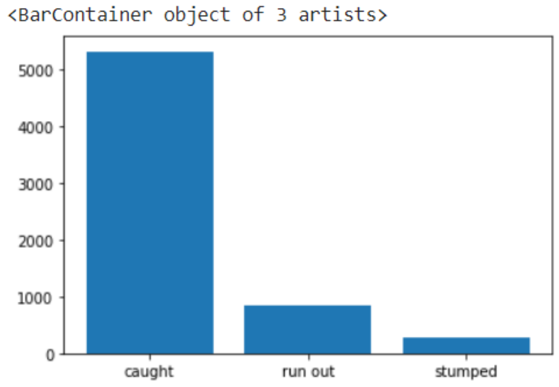
Figure 6. shows the most stumped dismissal analysis of a player in the history of IPL

The drawback in the figure 4 and 6 are that this depends on the number of matches a certain player has been chosen to play in. For example, MS Dhoni being one of the players who has participated in the most matches among the 10 players named in the figures.



	bat	field
Rajiv Gandhi International Stadium, Uppal	0.21	0.42
Maharashtra Cricket Association Stadium	0.50	0.68
Saurashtra Cricket Association Stadium	0.00	0.57
Holkar Cricket Stadium	0.00	0.88
M Chinnaswamy Stadium	0.44	0.55
Wankhede Stadium	0.50	0.51
Eden Gardens	0.43	0.63
Feroz Shah Kotla	0.47	0.55
Punjab Cricket Association IS Bindra Stadium, Mohali	0.35	0.54
Green Park	NaN	1.00
Sawai Mansingh Stadium	0.32	0.68
MA Chidambaram Stadium, Chepauk	0.61	0.38
Dr DY Patil Sports Academy	0.43	0.60
Newlands	0.75	0.33
St George's Park	0.43	NaN
Kingsmead	0.60	0.60
SuperSport Park	0.50	0.83
Buffalo Park	0.67	NaN
New Wanderers Stadium	0.00	0.50

Figure 2. Toss Analysis for win



```
<BarContainer object of 3 artists>
```

Figure 3. Dismissal Analysis

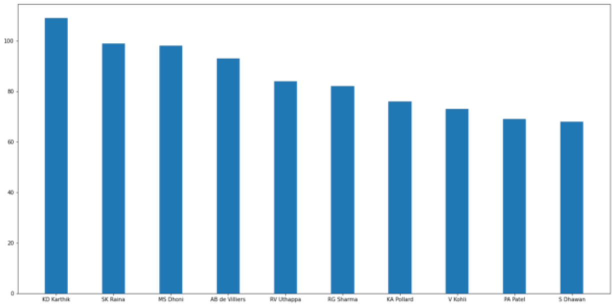
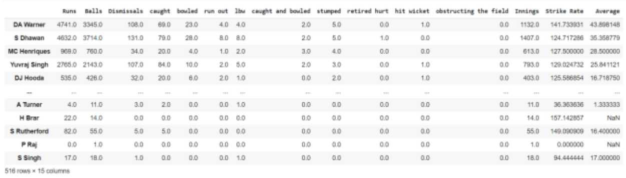


Figure 4. Most Caught Dismissal analysis of 10 batsmen



	Runs	Balls	Dismissals	caught	bowled	run out	lbw	caught and bowled	stumped	retired hurt	hit wicket	obstructing the field	Innings	Strike Rate	Average
DA Warner	4741.0	3345.0	108.0	69.0	23.0	4.0	4.0	2.0	5.0	0.0	1.0	0.0	1132.0	141.733931	43.898148
S Dhawan	4632.0	3714.0	131.0	79.0	28.0	8.0	8.0	2.0	5.0	1.0	0.0	0.0	1407.0	124.717286	35.358779
MC Henriques	969.0	760.0	34.0	20.0	4.0	1.0	2.0	3.0	4.0	0.0	0.0	0.0	613.0	127.500000	28.500000
Yuvraj Singh	2765.0	2143.0	107.0	84.0	10.0	2.0	5.0	2.0	3.0	0.0	1.0	0.0	793.0	129.024732	25.841121
DJ Hooda	535.0	426.0	32.0	20.0	6.0	2.0	1.0	0.0	2.0	0.0	1.0	0.0	403.0	125.586854	16.718750
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
A Turner	4.0	11.0	3.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	11.0	36.363636	1.333333
H Brar	22.0	14.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	157.142857	NaN
S Rutherford	82.0	55.0	5.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	55.0	149.090909	16.400000
P Raj	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.000000	NaN
S Singh	17.0	18.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	18.0	94.444444	17.000000

516 rows × 15 columns

Figure 5 shows the batting record of every batsman in IPL

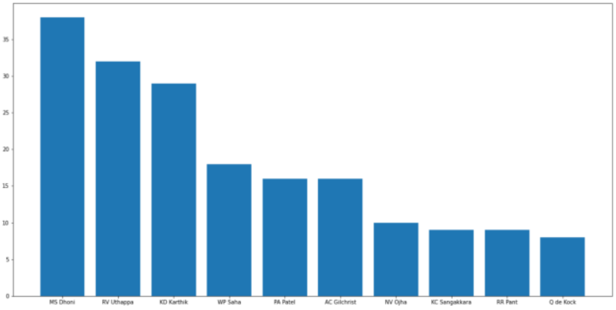


Figure 6. Most Stumped Dismissal analysis of 10 batsmen

Figure 5. was calculated based on the new dataset created out of the files obtained from the dataset on Kaggle.

V. PROPOSED WORK

We propose a new model that Rates all players and then given a playing a squad, providing a playing 11 out of it based on strength. To choose the model that suits our problem statement the best, we explore various machine learning algorithms. To predict the performance of players as how many runs will each batsman score and how many wickets will each bowler take for both the teams. Both the problems are targeted as classification problems where number of runs and number of wickets are classified in different ranges. We will explore naïve bayes, random forest, multiclass SVM and decision tree classifiers to generate the prediction models for both the problems. Random Forest classifier and KNN algorithms can be used to model the players ranking system.

The features that consist our model are: matches played, batting average, bowling average, type of player (batsman, bowler, all-rounder), strike rate, 3+ wicket hauls, 5+ wicket hauls, bowling economy, venue where the match is being played on and its respective toss win match win ratio (probability), overall current season performance of batsman and bowler, fielding – number of catches taken and run outs assisted, etc.

Another model that can be explored is that of Factor Analysis. Factor Analysis on which the features such as performances of batting, bowling and fielding depends on respectively are taken. For each of the department, the features can be brought down to factors and analyze the relative important of these factors that affect batting, bowling and fielding performance. These factors can be used along with other derived metrics to individually rate the players based on how well they perform on these department specific factors.

REFERENCES

[1] Vaibhav Khatavkar, Parag Kulkarni. “Context Based Cricket Player Evaluation Using Statistical Analysis”. In, International Journal of Knowledge Based Computer Systems 7 (1), June 2019, 01-0

[2] Madan Gopal Jhawar, Vikram Pudi. “Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach”. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2016) Conference Center, Riva del Garda.

[3] Hemanta Saikia, Dibyojyoti Bhattacharjee “On Classification of All-rounders of the Indian Premier League (IPL): A Bayesian Approach”. Vikalpa: The Journal for Decision Makers, Oct 2011.