Optimal IPL Playing 11 Team Selection

Github link: NihalChengappa/Batch-13_Optimal-IPL-team (github.com)

Group members:

1. Nihal Chengappa P.A PES2UG20CS224
2. Nilesh Ravichandran PES2UG20CS225
3. Pranav Rao Rebala PES2UG20CS248

Abstract: The Indian Premier League or its acronym-IPL is a worldwide famous and prestigious T20 cricket competition where the best cricketers all over the world participate. This paper centers on recommending the best possible Playing XI from given a set of players. The model here predicts the best possible set of batsmen and bowlers by predicting their scores and economy respectively in the next match. The model uses Regression Analysis and hence gives low error.

Keywords—IPL, Cricket, Regression Analysis, Economy, Playing XI

I. INTRODUCTION

The IPL first came into existence in 2008 and is presently in its 13th season, the BCCI being the crux for the start of the IPL and its prestige too. It is held every year across India and during this time, almost no other international cricket takes place, enabling the players across the globe to feature in this league. The victory of the IPL can be ascribed to a number of things but a significant reason to it is the quality of cricket it produced, consequently giving quality amusement for the group of onlookers/viewers. It is presently considered to be the leading T20 competition in the world. The franchises who own the teams take part in auctions every year with a pool of players and not just the international but domestic players go into the auction as well. Franchises are permitted to have a maximum of 25 players with max of 10 overseas players in the squad (only max of 4 can feature in the playing 11) after the auction and can release any number of players from their squad but can retain only a specific number (differs every year) before the next auction and the released players can feature in the following year's auction along with the new entries (new pool of players).

A lot of homework is done by the franchises- a lot of money is spent on buying the players and so a ton of analysis is done and from the squad, analysis is again done to choose a playing 11. The T20 format is an unconventional one where the game is considered to be more in the batsmen's favor with the power play consisting of the first 6 overs with only 2 fielders allowed outside the 30-yard circle and then the final five overs being the death overs and hence the teams provide more accentuation on the quality of their batting line up. The teams would generally prefer to have a mix of right and left handedness to unsettle the bowling team. Versatility is a boon in this format, here the top 3 generally consist of one who carries the innings and the rest play with a higher strike rate. Then followed by the middle order who form the crux of the batting department are known to provide the necessary impetus and take the game till the end. These consist of the power hitters mainly. Players who can bat and bowl-all-rounders that is, are vital. They provide the necessary balance in the team. The lower order is just expected to swing their bat. The analysis is also done for the bowlers since they are just as vital as anyone else in the team. Good Spinners and genuine pace bowlers are quintessential- Swing/Pace bowlers with the new ball upfront and then spinners to control the middle overs followed by death over specialists are a key. The control of economy is the key in this analysis for bowlers since lesser the runs given by the bowlers, better is the result. An analysis for all-rounders is also done in such a way where players with data who have batted and bowled are taken in the list with the analyst comparing the statistics of the batting and bowling and choosing whom with a better overall record would fit in. The playing 11 overall is chosen in such a way where a max of 4 foreigners can be picked.

All this has a lot of time and thinking involved, lot of analysis is done on each and every player and then the decision to whether or not would the player be selected is decided based on his past and expected performance. In this model, we use Data Analytics techniques to prepare the model. Exploratory Data Analysis (EDA) is done on the deliveries dataset to get the required understanding and information of the dataset. We then will be able to decide on how to go about the analysis and what techniques to use to build an ideal model and predict the best possible set of players that should be chosen from the given squad.

II. RELATED WORK

A lot of measurable investigation has been done over the recent decade on cricket. There has been substantial increase in the need of performing investigation on the Indian Premier League due to the huge amount of cash contributed in it. This had led to analysts taking the route of predicting match results, predicting auction costs of different cricketers etc. The particular IPL teams' administration to incorporate of data analysts to supply input to the team's proprietors and coach on techniques on which player to select and for what cost. They too play a critical role in proposing changes within the different areas where performance is frail such as fielding, a batsman's strike rate, a bowler's economy, player's health based on his food intake and his training regime.

[1] have concocted a strategy that ranks players based on the following factor - Deep Performance Index method to evaluate a player. The algorithm considers setting of the player in terms of lists like most valuable player index for batsman and for bowler. The critical setting presented is the winning contribution ratio of the player. An approach for assessing a player based on the venue of match, by and large

normal, current season normal and extra parameters.

Batting Rating Points System = Winning Contribution Ratio + Average Venue Index + (Batsman Current Season Average / Batsman Overall Average * Batsman Current Season Strike Rate / Batsman Overall Strike Rate) + HPR

Bowling Rating Points System = Winning Contribution Ratio + Average Venue Index + (Bowler Overall Average / Bowler Current Season Average * Bowler Current Season Average / Bowler Current Season Economy) + Wickets Matches Ratio

The values of BRPS (Batting Rating Points System) varies in the range of (3.47 - 8.42) and that of BoRPS (Bowling Rating Points System) varies in the range of (2.58 - 6.54). Although, to apply the above calculations the taking after presumptions were necessary: Factors like team support/assists were ignored (Team Contribution). [1] Only considered IPL performance data up to IPL 7 and the generally T20 career information up to the conclusion of IPL 8 of all the players taking an interest in IPL. Batsman scoring more than 500 runs in twenty-20 internationals with a strike rate surpassing 100 and who have played in at slightest 25 matches were considered for batting ranking. Applying these criteria, 89 batsmen were considered for this rating system.

Barr and Kantor et al. [4] utilized portfolio examination to decide the set of batsmen who are gathered to be more appropriate for a given one-day squad.

Gerber and Sharp et al. [5] proposed an integer programming strategy for selecting a limited over squad or 15 players instead of playing 11 players. The method included to gather information from 32 South African players to choose the ODI squad of 15 players.

[2] provides insights about modelling batsmen, bowlers and eventually modelling groups (rating) and proposing a 'Playing XI', given a squad/pool of players employing a few features. [2] propose novel strategies to demonstrate batsmen, bowlers and teams, utilizing different career insights and recent performances of the players. They moreover propose that their model can foresee the winner of ODI cricket matches, by employing a novel dynamic approach to reflect changes in player combinations. A parcel of assumptions have been made within the modelling step of batsmen, bowlers and the teams as a whole.

[2] use features such as Matches Played, Batting Innings, Batting Average, Number of Centuries, Number of Fifties, Bowling Innings, 5Wkt Hauls, Bowling Average and Bowling Economy. Numerous imperative highlights such as venue conditions and relative quality within the team have been disregarded. Another critical concept is that all-rounders have not been considered independently but they are taken as batsmen or bowlers. Thus, our paper points to too show all-rounders to boost the playing XI prediction quality. Our model provides more value to recent performances as the players learn and grow.
Although [2] has been executed for ODI cricket matches, IPL T20 would be a more challenging task and consequently would be more important in terms of need in today's world. [2] claim that the k-Nearest Neighbor (kNN) algorithm utilized gives way better comes about with an exactness of 71% as compared to other classifiers models that have an exactness of 56% and 63%

[3] provides understanding on how to handle modelling all-rounders which was not taken care of in [2]. The paper takes into consideration that the player was within the playing eleven for at least 5 matches within the IPL, the player has bowled for at slightest 10 overs in IPL, the player has faced at slightest 100 balls within the IPL, and, as it were a add up to of 35 players have been considered for the investigate. Subsequently, it does not speak to the entire populace 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 determining the anticipated class of all-rounders based on the significant predictors can be a method to approach our problem. [3] based on the above strategy have claimed that all-rounders have been effectively classified into 4 non-overlapping categories, to be specific- Performing all-rounders, batting all- rounders, bowling all-rounders and Under-performer with a precision of their demonstrate at 66.7%.

III. PROPOSED SOLUTION

In order to select the best playing 11 for an IPL team we propose a solution involving selecting batsman and bowlers in the ratio of m:n (depends on how many batsman and bowlers a team wishes to have) we will be using the data we have accumulated over the performances over the past decade(specifically 2008-2019).
We will need 2 models one for modelling the best batsmen and another for modelling the best bowlers. But before building the models we will need to perform some preprocessing-

A. Pre-processing

The file deliveries.csv has at least 431000 null values in the attribute's fields: player_dismissed, dismissal_kind and fielder. All the empty places were replaced with NaN to avoid errors and make the dataset consistent overall.
The columns bye_runs,legbye_runs,is_super_over are not usefull for our predictions and hence can be dropped.
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. However since no of innings doesn't contribute to our predictions that column can be dropped as well.

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.

B. FOR BATSMEN:

Model Building:
The first model we need to build is for predicting the best batsmans.This model predicts the runs scored by a batsman by taking the following features-

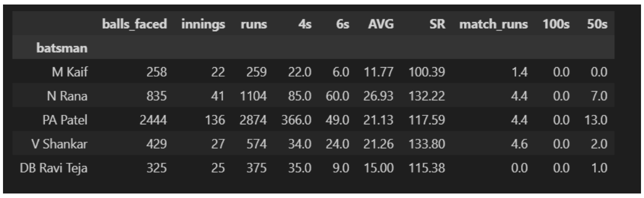
a. Balls faced
b. No of innings played
c. Total runs scored
d. Total 4s
e. Total 6s
f. Average score per game
g. Strike rate
h. Total 100s
i. Total 50s

To do this first we create a dataframe with the summary statistics of each batsman in our dataset having the columns *balls faced, no of innings played, total runs scored,4s,6s,average runs, strike rate, total 100s,total 50s and match_runs which is the average runs scored by the batsman in the last 5 matches.*

While computing the total innings played by a batsman we take values greater than the mean of innings played by all the batsmen as some batmen who come down the order may play fewer matches hence will not have a very accurate average score

Algorithm:
batsmandf=batsmandf[‘innings’]>batsmandf[‘innings’].
mean()]
#computes the batsmen who have played greater than
the mean number of innings

Our final batsman dataframe would look like this-



	balls_faced	innings	runs	4s	6s	AVG	SR	match_runs	100s	50s
batsman										
M Kaif	258	22	259	22.0	6.0	11.77	100.39	1.4	0.0	0.0
N Rana	835	41	1104	85.0	60.0	26.93	132.22	4.4	0.0	7.0
PA Patel	2444	136	2874	366.0	49.0	21.13	117.59	4.4	0.0	13.0
V Shankar	429	27	574	34.0	24.0	21.26	133.80	4.6	0.0	2.0
DB Ravi Teja	325	25	375	35.0	9.0	15.00	115.38	0.0	0.0	1.0

Fig III.1

Now we take our match_runs variable as the target and split the remaining features into test and train datasets in the ratio of 0.8:0.2.

Now we must build our models:

1. Linear Regression :

   Features: *balls faced, no of innings, total runs ,4s,6s,average runs, strike rate, 100s, 50s*
   Target: Match_runs(average runs scored in last 5 matches)

   This model is then tested against our test data and its accuracy is calculated using RMSE

   This mode has an RMSE score of 3.65 which is pretty good thus proving that this model is giving us accurate predictions.

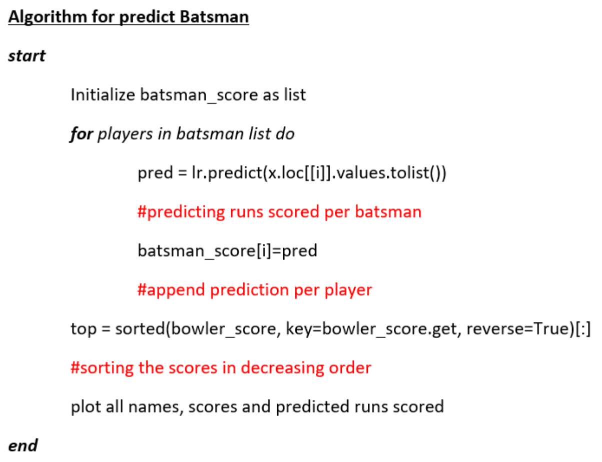
2. Random Forest :

   Features: *balls faced, no of innings, total runs ,4s,6s,average runs, strike rate, 100s, 50s*
   Target: Match_runs(average runs scored in last 5 matches)
   This model is then tested against our test data and its accuracy is calculated using RMSE

   This mode has an RMSE score of 6.91 which is pretty good however linear regression is giving better results.

Thus we will be using LR model for our predictions

Predictions:



```
Algorithm for predict Batsman
start
        Initialize batsman_score as list
        for players in batsman list do
                pred = lr.predict(x.loc[[i]].values.tolist())
                #predicting runs scored per batsman
                batsman_score[i]=pred
                #append prediction per player
        top = sorted(bowler_score, key=bowler_score.get, reverse=True)[:]
        #sorting the scores in decreasing order
        plot all names, scores and predicted runs scored
end
```

Fig III.2

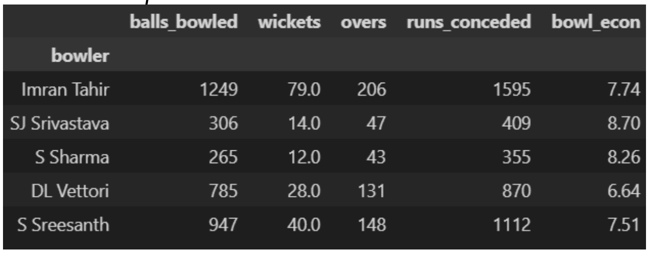
To predict the top N batsmen we created a prediction function using algorithm defined in Fig III.2. This function as its argument takes a list of N different batsmen whom we wish to compare. It first runs a for loop which predicts the runs scored for the N batsman that we pass as arguments to it. Then we can graphically compare our predictions and decide which batmen to pick for our team.

C. *FOR BOWLERS:*

Model Building:
The second model we need to make is the bowlers model. This model should be able to predict the economy of a given bowler accurately. Lower the economy the better. The features of this model include-

a. *Balls bowled*
b. *Wickets*
c. *Overs*
d. *Runs conceded*
e. *Bowlers economy*

Bowlers dataframe:



	balls_bowled	wickets	overs	runs_conceded	bowl_econ
bowler					
Imran Tahir	1249	79.0	206	1595	7.74
SJ Srivastava	306	14.0	47	409	8.70
S Sharma	265	12.0	43	355	8.26
DL Vettori	785	28.0	131	870	6.64
S Sreesanth	947	40.0	148	1112	7.51

Fig III.3

After this we calculate the total runs conceded by the bowler in the last 5 matches. Using this we calculate the bowler’s recent economy which serves as the target variable for our model.
Now we take our recent economy variable as the target and split the remaining features into test and train datasets in the ratio of 0.8:0.2.

Now we must build our models:

1. Linear Regression :

   Features: *balls bowled, wickets, overs , runs conceded, bowler economy*

Target: Recent economy (average economy of bowler in last 5 matches)

This model is then tested against our test data and its accurancy is calculated using rmse value

We find that the rmse value is 1.8 which proves that our model is giving pretty accurate predictions.

2. Random Forest :

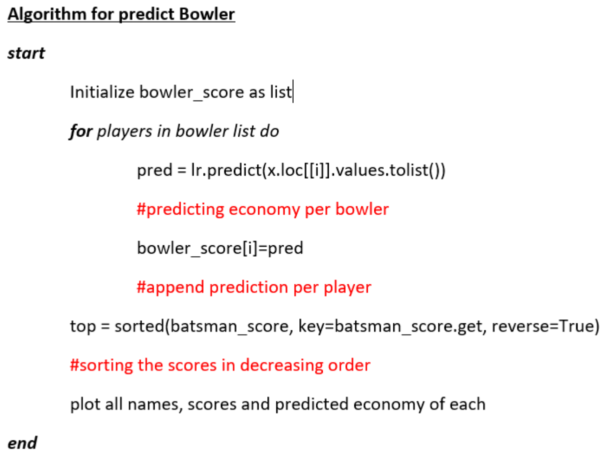
   Features: *balls bowled, wickets, overs , runs conceded, bowler economy*
   Target: Recent economy (average economy of bowler in last 5 matches)

   This model is then tested against our test data and its accuracy is calculated using RMSE

   This mode has an RMSE score of 1.904 which is pretty good however linear regression is giving better results.

   Thus we will be using LR model for our predictions

Predictions:



```
Algorithm for predict Bowler
start
        Initialize bowler_score as list
        for players in bowler list do
                pred = lr.predict(x.loc[[i]].values.tolist())
                #predicting economy per bowler
                bowler_score[i]=pred
                #append prediction per player
        top = sorted(batsman_score, key=batsman_score.get, reverse=True)
        #sorting the scores in decreasing order
        plot all names, scores and predicted economy of each
end
```

Fig III.4

To predict the top N bowlers we created a prediction function using algorithm defined in Fig III.4. This function as its argument takes a list of N different bowlers whom we wish to compare. It first runs a for loop which predicts the economy for the N bowlers that we pass as arguments to it. Then we can graphically compare our predictions and decide which bowlers to pick for our team.

IV. Experimental Results

Through our data analysis and models we have figured out how to predict the top batsmen and bowlers for a team. Now lets try to experiment and build a team of 5 batsmen,4 bowlers and 2 all-rounders. For this we first need to know the players in the auction pool.

Auction pool:
SK Raina,V Kohli,RV Uthappa,Yuvraj Singh,PA Patel,R Dravid,G Gambhir,KD Karthik,RG Sharma,V Sehwag,SR Watson, P Kumar,R Ashwin,IK Pathan,A Mishra,AB Agarkar, AB Dinda,AD Mathews, AD Russell, B Kumar

Now we put all players who can bat in the predict_batsman function:
SK Raina, V Kohli, RV Uthappa, Yuvraj Singh, PA Patel, R Dravid, G Gambhir, KD Karthik, RG Sharma, V Sehwag, SR Watson, IK Pathan

Now we put all players who can bat in the predict_bowler function:
SK Raina, Yuvraj Singh, SR Watson , P Kumar, R Ashwin, IK Pathan, A Mishra, AB Agarkar, AB Dinda, AD Mathews, AD Russell, B Kumar

On finishing our predictions we get the following graphs:

Batsman:

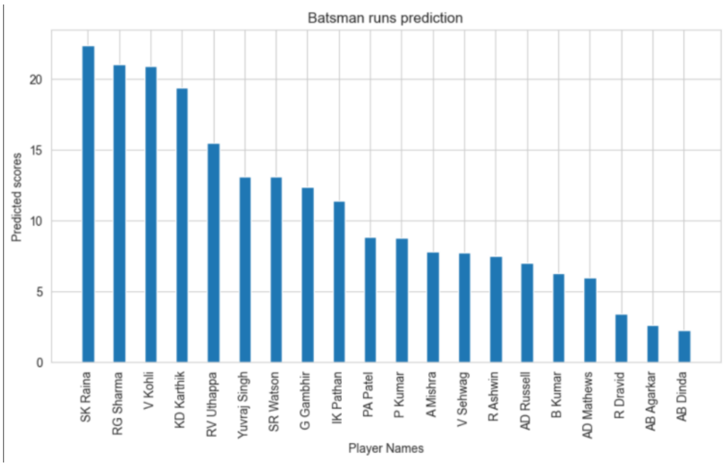


Fig IV.1

Bowlers:

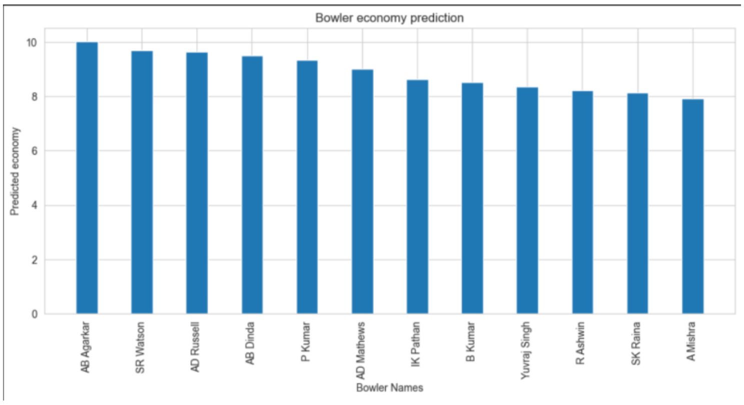


Fig IV.2

On analyzing the graphs (Fig IV.1 and Fig IV.2) we come to the following conclusions:

a. The top 2 all-rounders : SK Raina, Yuvraj Singh
b. The top 5 batsmen are:
   RG Sharma, KD Karthik, V Kohli, RV Uthappa, SR Watson
c. The top 4 bowlers are:
   A Mishra, R Ashwin, B Kumar, IK Pathan.

This is not a fixed format and as per the requirements we are free to chose x, y and z number of batsmen, bowlers and all-rounders respectively.
Thus we have chosen the optimal 11 players for our IPL

team. Not only are we able to account for batsmen and bowlers but we can also predict required number of all-rounders by carefully analyzing our graphs.

V. CONCLUSIONS

After comparing the RMSE (Root Mean Squared Values) values from the two implemented models, Linear Regression yields the best result followed by the Random Forest technique. The model is present in the range of analyzation of data and prediction of performance of players with the conclusion of our model being- it can serve as a recommendation on a real-world level and can propose the best possible playing eleven for the team, it predicts how much a batsman would score in the next match taking into consideration a number of factors like runs, strike rate, average, balls faced, boundaries, 50s, 100s and innings and for the bowlers on the other hand, it predicts their expected economy in the next match which is pivotal through factors such as balls bowled, wickets taken, runs conceded and economy from the training data because lesser the runs conceded, better it is for the team. Similarly, it is done for all-rounders too where we feed in data of players who have batted and bowled as well and choose the best set of the players with a good overall record and hence strengthen the balance of the team giving more depth in both departments.

The real advantage of this model is that we do not have to select a fixed number of batsmen, bowlers or all-rounders for a particular match, i.e., the team can decide what combination to play, i.e., how many batsmen, all-rounders and bowlers to play depending on the conditions- we can play 3 bowlers and 5 batsmen with 3 all-rounders if it's a bowling pitch or 4 batsmen and 5 bowlers with 2 all-rounders if it's a batting pitch, etc accordingly. Another advantage is that our model can take a specific list of players which the team wants, it doesn't have to take all the players from the training data, so if we want to select 2 batsmen from a list of 10 then we can feed 10 players' statistics and predict who would score more in the next match and select them accordingly and similarly for the bowlers etc. The novelty in the model is that we have predicted the players to be selected in a different way when compared to certain other models by predicting their future performance. Hence, satisfactory prediction results can be expected to help the team.

VI. FUTURE WORK

There is never an end to modifying and bettering the model. The future work for our project would be to elevate this to the next stage- like picking a wicketkeeper, openers, all-rounders, power hitters, spinners and pacers on its own rather than us providing a list and even take the venue into consideration as well.

VII.CONTRIBUTION OF TEAM MEMBERS

1. Nihal Chengappa P.A (PES2UG20CS224)
   - Analysis, Model Fitting, Testing and Prediction for bowlers
   - Proposed Solution, Experimental results of the Report

2. Nilesh Ravichandran (PES2UG20CS225)
   - Analysis, Model Fitting, Testing and Prediction for batsmen
   - Introduction and Conclusions with Future work of the Report

3. Pranav Rao Rebala (PES2UG20CS248)
   - Data cleaning and Exploratory Data Analysis
   - Related Work and Appendix of the Report

VIII. APPENDIX

The appendix section basically consists of the Exploratory Data Analysis and data visualization. The EDA is the base for any model that helps us to analyze and comprehend the different features in the data and hence helps the analyst to accordingly decide on how to go about the model and which features would be of greater importance.

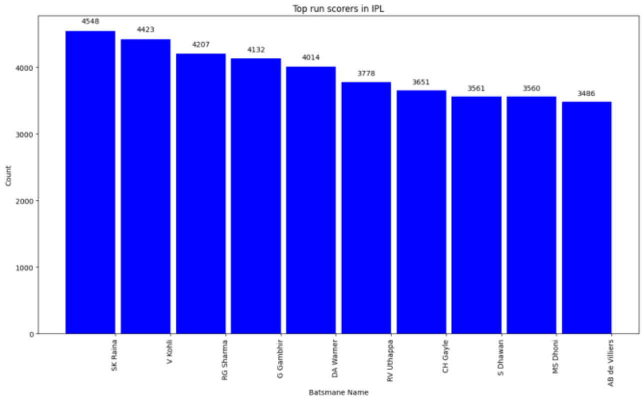


Fig VIII.1

Fig VIII.1 shows the top 10 run scorers in the IPL with Raina leading the list

Although the list in the figure above consists of some good players, it could be a drawback where some good players don’t make the cut in this list because of the fact that they might have played few matches.

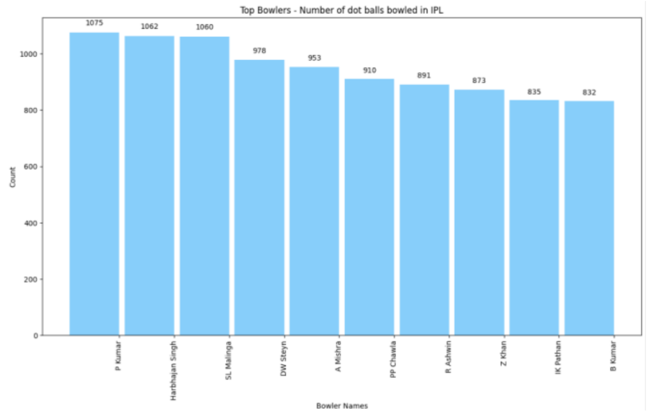


Fig VIII.2

Fig VIII.2 shows the top 10 bowlers in the IPL who have bowled the most number of dot balls

Just like V.1, this too consists of the good players but some players with lesser matches might not feature in this list due to less number of overs under their belt.

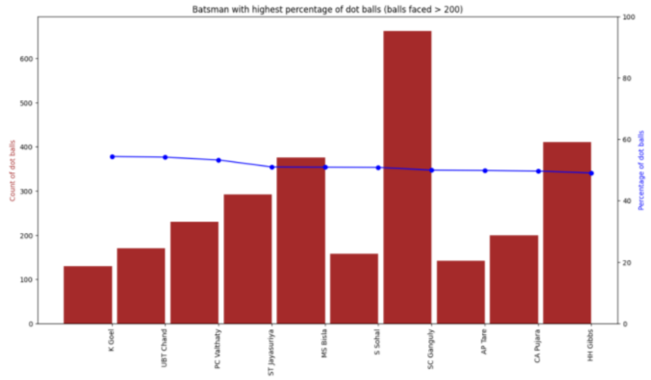


Fig VIII.3

Fig VIII.3 shows the batsmen with the highest percentage of dot balls after facing a minimum of 200 deliveries

Fig VIII.3 has really good analyzation to a great extent but the drawback being that there could be players in this list who accumulate a lot of dot balls yet score a lot due to scoring a number of boundaries albeit a rare instance but can exist.

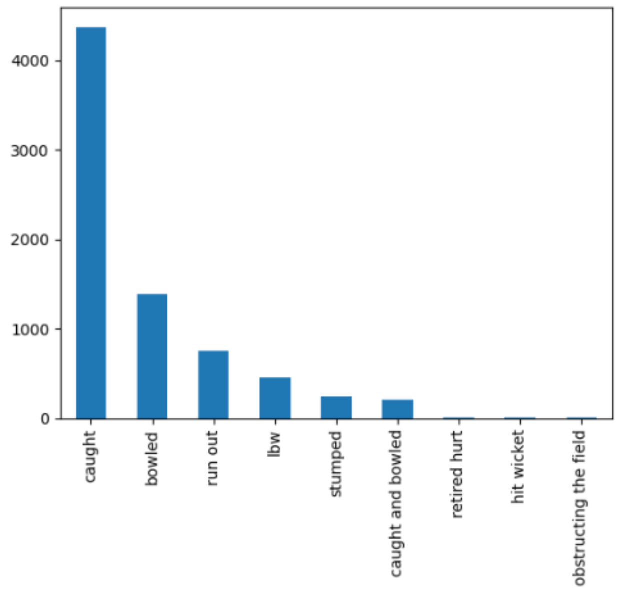


Fig. VIII.4

Fig VIII.4 shows the different modes of dismissals and majority of them are 'caught'

A lot can be inferred from Fig VIII.4 and hence we can accordingly go about our analysis

IX. 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.

[4] Barr, G. D. I. and Kantor, B. S. (2004), A criterion for comparing and selecting batsmen in limited overs cricket, Journal of the Operational Research Society, 55(12), 1266-1274.

[5] Gerber, H. and Sharp, G. D. (2006), Selected a limited overs format of cricket squad. They choose using an integer programming model, South African Journal for Research in Sport, Physical Education and Recreation, 28(2), 81-90.