International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Proposing Competent Team Composition for T20 Cricket Through Data Processing Techniques

Madisetty Eshwar, Yasvantha Sai Atmuri, Saiteja Kalam
Department of ECE, Amrita Vishwa Vidyapeetham, India

Article Info
Volume 9, Issue 5
Page Number : 563-570

Publication Issue
September-October-2022

Article History
Accepted : 10 Oct 2022
Published : 28 Oct 2022

ABSTRACT

In cricket, selecting the best playing XI is an important task to win the game. There are various factors such as venue, performance of batsman and bowler; and opposition that influence the team selection. This project predicts different parameters such as run-rate, strike rate, economy and wickets, and use them for construction of an ideal team; these predictions are done by using various machine learning algorithms namely K nearest neighbor, random forest and gradient boosting which uses the past data for prediction. Different roles require different skill-set from the players. This paper selects players based on the predicted values and suggests a suitable playing XI.

Keywords— Cricket, Twenty-Twenty (T20), Team Prediction, Data Processing, Sports Analytics

I. INTRODUCTION

The business of Sports or Sports market has been receiving increased attention with the enhancement in technologies to telecast sports events to every corner of the globe. From the conventional Olympics, World Cups, Continental Competitions, to the present day Premier Leagues, the sports market contributes heavily to the global economy. The initial origin of sports business could be traced to brand endorsements by leading sports personalities. The evolution and popularity of the Premier Leagues and the mass "fan following" commanded by teams have completely altered the Sports market scenario. Football clubs including Real Madrid and Manchester United are as popular in Asia and every other part of the world as they are in Europe. Cricket is also catching up with its own version of premier leagues and club franchisees. Brand endorsement by sports clubs of primarily sports merchandise is a good source of income for the clubs. This endorsement is fast catching up in soft drinks and fashion wear. Anand and Arjun [1] provide a detailed study of impact of fan behaviour on sales of branded sports merchandise. The digital broadcast technology associated with sports event telecast generates immense data, attractive for employing machine learning and data analytics techniques, providing scope for evolving the existing sports market into a larger Sports Industry [2]. An emerging allied sector is the ESports domain where technological interventions have already made big headway including assisting game participants evolve techniques, predict the winners [3].

Sporting events involving football and cricket dominate the sports market scenario at present. Both the games involve extensive decision making by the on-field referees or the umpires. The decisions heavily impact the flow of the game and a wrong decision could be hurting teams, fans and viewers. There is an indirect impact on the revenue generation component.

Cricket has three game formats, at international level including Test (Five day event), One Day Match and a shortened Twenty-Twenty (T20) format. This provides scope for more business avenues, as evinced by the success of the Premier League competitions involving the T20 format. Incorporating technological interventions to assist the umpires in cricket has been an upcoming field, with the advent of technologies including Hawk –eye, Snickometer, Hotspot. Academic research on technological interventions in sports has been steadily increasing in the last decade. Automated decision making for run outs employing the frames captured for telecast has been explored by Sabarish et al [4]. Established algorithms including Edge detection and Absolute difference were employed. In the same line, Ganesh et al [5] report automated decision making in cricket employing SVM and CNN techniques. Data mining and classifier techniques to predict the winner mid-way through the match has been reported by Tejinder et al [6]. The techniques mentioned could be classified as decision making employing "within-match" data.

Another set of techniques correspond to decision making employing "pre-match" data. Predicting a player's performance in future matches, based on his past performances has been explored in [7]. The prediction has been done for the One Day International format. In this paper, we propose novel techniques employing established data processing algorithms to propose a competent team or "Playing XI" as it is referred to in the sport domain, for the Twenty-Twenty (T20) format.

II. METHODOLOGY

A. Data set details

Previously available data is used to predict different parameters. This data is used to make decisions regarding a player. The first step is to collect the data. The data set is collected from [1] which consists of T20 International matches played by India from 2010-2020. There are a total of three data sets which are as follows.

1. Data set-1: India T20I, This data set consists of ball- to-ball details of 132 T20 International matches played by India from 2010-2020.
2. Data set-2: Match details – This data set contains the details of the match.
3. Data set-3: Player details – This data set contains the data of the entire Indian players till 2020.

B. *Machine learning algorithms employed*

1. K-Nearest Neighbor Regression
2. Random Forest Algorithm
3. Gradient Boosting Algorithm
4. Linear Regression
5. Decision Tree Algorithm

C. *Cricketing Parameters for data analysis*

1. Batting Average
2. Batsman Strike rate
3. Run rate of the team
4. Economy of the bowler
5. Bowling average
6. Wickets taken by a bowler

D. *Mathematical parameters*

1. Mean absolute error

Mean Absolute Error (MAE)is a measure of errors between paired observations expressing the same phenomenon.

2. Mean absolute percentage error

Mean Absolute Percentage Error (MAPE) is a statistical measure of how accurate a forecast system is.

3. Accuracy

One metric for assessing classification models is accuracy. Informally, accuracy refers to the percentage of correct predictions made by our model.

4. GINI

Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly.

For the prediction of cricketing parameters, each algorithm undergoes a certain process which gives the predicted output. Every algorithm used in this project undergoes the same process.

Run rate of the team

For predicting run-rate of the team, the input parameters considered are opposition team, venue of the match and innings which the team is playing. Mean absolute percentage errors are calibrated after evaluating with different algorithms shown in **Table 1.** Since KNN regression yields the lowest error percentage, KNN regression performs well for this model.

To show the working of the algorithm let us assume the following:

Innings = 1, Opposition team = Australia, Venue = Seddon Park

KNN regression calculates the average run rate of all 5 nearest points (Here, the value of K=5) which is the predicted run rate of the input point i.e. 7.86 runs per over. In order to represent the run rate in the form of range, mean absolute error (0.61) is added and subtracted to the average and represented as [7.25,8.49].

Table 1: Error percentages of various algorithms for the parameter—run rate of the team.

Algorithm Name	Mean Absolute Percentage Error
KNN Regression	15.86%
Gradient Boosting	16.20%
Linear Regression	16.45%
Random Forest	18.58%
Decision Trees	26.21%

Economy of the bowler in respective over-type

From Table 2, as KNN regression produced the least error percentage from the selected algorithms, economy of the bowler in respective over-type was done using KNN regression. Criteria's considered are opposition team, Innings played, Venue, over-type and name of the bowler.

KNN regression uses the Euclidean distance formula to calculate the 5 nearest points from the input. Economy of the bowler in respective over-type for given criteria can be calculated as follows; let's assume the given criteria as follows:

Innings = 1, Opposition team = England, Venue = Kings mead, Over-type = Death, Bowler = JJ Bumrah

Applying the average formula, KNN regression calculates the average of economies of all the 5 nearest points which was found with the Euclidean distance formula which is 7.7. Mean absolute error is added and subtracted to the output to represent in the form of range i.e. [6.28, 9.12].

Table 2: Error percentages of various algorithms for the parameter – Economy of the bowler in respective over-type.

Algorithm Name	Mean Absolute Percentage Error
KNN Regression	48.82%
Gradient Boosting	58.48%
Linear Regression	61.90%
Random Forest	56.48%
Decision Trees	68.92%

Economy of the bowler in the whole match

Here, it is the economy of all the balls that the bowler has bowled throughout the match. The criterion considered for calibrating the economy are opposition team, venue and name of the bowler. KNN regression produces least mean absolute percentage error when compared to other algorithms listed in **Table 3**. Hence, KNN regression algorithm is used.

To illustrate the working of the algorithm, input parameters considered are as follows:

Opposition team = England, Venue = Seddon Park, Bowler

= Bhuvaneshwar Kumar

With the help of the Euclidean distance formula, 5 nearest indices are marked and the average of their economies is calculated and the output is 5.78 for this example.

Table 3 : Error percentages of various algorithms for the parameter – Economy of the bowler for the whole match.

Algorithm Name	Mean Absolute Percentage Error
KNN Regression	34.01%
Gradient Boosting	34.52%
Linear Regression	34.61%
Random Forest	36.83%
Decision Trees	47.99%

Wickets taken by a bowler in respective over-type

Classifier algorithms must be implemented instead of regression algorithms for this model as classifiers are good at handling integer values. Wickets taken by the bowler are integer values, hence the classifier algorithm is used. Parameters involved in predicting the wickets are name of the bowler, Opposition team, Innings, venue and over-type. From Table 4, we can see that KNN and gradient have high accuracies yet random forest is most preferred due to imbalance class behaviour. It is the phenomenon when there is a single value repeated more than half of the list length the algorithm gives the most repeated output without learning.

To explain the working of the algorithm, the sample inputs are as follows:

Innings = 1, Opposition team = Australia, Venue = Seddon Park, Over-type = Death, Bowler = JJ Bumrah

The algorithm forms a decision path as given below

[0, 977, 1998, 3035,3964,4909, 5880, 6799, 7820, 8771, 9760, 10737, 11698, 12629, 13584, 14597, 15658, 16639, 17570, 18515, 19464].

The algorithm starts from 0 and moves to the next location with its nearest values for 20 times as n_estimators=20. The average of all the outputs is calculated. In this criteria mentioned above, the number of wickets taken by JJ Bumrah in death overs is 0.

Table 4: Accuracies of different classifiers for wickets taken by the bowler at respective over-type

Algorithm Name	Accuracies
KNN Classifier	74.9%
Gradient Boosting Classifier	74.9%
Random Forest Classifier	71.1%
Decision Tree Classifier	63.5%

Strike rate of the batsman in respective over-type

It is the average runs scored per 100 balls faced by the batsman. The parameters contributing for the prediction are name of the batsman, Venue, innings, opposition team name and over-type.

The data associated with strike rate are trained with several algorithms listed in Table 5. It is found that gradient boosting has least mean absolute percentage error and hence this algorithm is used for this model. To explain the working of the algorithm, sample inputs are given as follows:

Innings = 1, Opposition team = Pakistan, Venue = Karachi, Batsman = RG Sharma, Over-type = Death
Gradient boosting algorithm takes the inputs and a decision tree is formed. The residual output is the input for the next decision tree for the next 100 decision trees.
Final predicted output : BV+
(LR*RP1)+(LR*RP2)+….(LR*Rpn)
Where, BV is base value, LR is the learning rate, RP is the residual of the nth decision tree and n=100. Now based on the residual values predicted, the strike rate for the above mentioned criteria is 110. In order to represent the run rate in the form of range, mean absolute error (17.63) is added and subtracted to the average and represented as [92.37,127.63].
Table 5 : Error percentages of various algorithms for the parameter – strike rate of the batsman in respective over-type

Algorithm	Mean absolute percentage error
KNN regression	51.72%
Gradient Boosting	46.32%
Linear Regression	47.10%
Random Forest	46.47%
Decision Trees	64.90%

III. RESULTS

Generally in any sport, each player will possess a different skill-set. Captain or team management need to assign different roles based on their skill-set which would result in balance of the team. A team is considered to be balanced when it is good at all aspects. To maintain the integrity and balance of the team, we split up the 11 players into 4 batsmen, 1 wicket-keeper, 2 all-rounders and 4 bowlers respectively.

The split-up that we mentioned was tentative. In case of weak performance or injuries, we should have other options to balance the team performance. For suppose, if a bowler has a bad day on the field, then the other bowling option will come into play. From the scenario considered above, this model is going to suggest the best playing XI using the above predicted parameters.

A. Batsmen categorization

Batting order is the order in which batsmen come to bat on the field with 2 batsmen on either side of the pitch i.e. strike and non-strike end. The order in which the eleven players bat is generally decided before the start of a cricket match and yet it can change over the course of the game.

The batting order is based upon several factors such as player abilities, degree of comfort, potential batting combinations and situation of the match. Hence, considering each factor, the batting order was classified into three categories in common parlance.

- Top Order (1-3 positions)
- Middle Order (4-7 positions)
- Upper Middle Order
- Lower Middle Order
- Lower Order (8-11 positions)

B. Bowler categorization

Bowling category consists of quick bowlers, whose primary weapon is pace, swing and groove bowlers, who aim to deviate the ball's path through the air are all examples of fast bowlers. A spin bowler throws the ball gently and spins it. When bouncing off the pitch, allowing it to turn at an angle. The spinners are generally categorized into two, finger spinners and wrist spinners.

C. Game Conditions

Several parameters are involved in predicting the most probable playing XI such as

- Opposition Team
- Venue
- Minimum experience required for a player (Number of matches a cricketer played).

Minimum experience shortlists the players who are eligible for the team. The shortlisted players will go through algorithms and specific conditions in order to give the most probable playing XI as the output.

D. Batsmen Selection

First step to predict the playing XI is selecting the top order. Top order includes three pure batsmen who score more runs quickly with consistency. To achieve this, a batsman with a high strike rate and good average are required. Using data scraping techniques, the data set is trimmed such that only top-order batsmen would iterate through a gradient **boosting algorithm** which predicts the strike rate of the batsmen at different stages of a match (Power play, Middle Over's and Death Overs). Average is computed of all three values and sorted accordingly and again batting average is taken into consideration and sorted again. First 3 players of the list are considered as top order batsmen.

A wicket keeper can always be a pure batsman or an all rounder. Our next step is to select a wicket keeper. There are mainly 2 aspects involved.

1. If there is a wicket-keeper in top order, then we need not include a wicket-keeper in the middle order.
2. If there is no wicket-keeper in top order, then we must include a wicket-keeper as one of the batsmen in the middle order.

For Middle order, again there are 2 categories, upper middle order and lower middle order. In this scenario, the average of the batsman is not considered because the main aspect for the middle order batsmen is the strike rate.

All the players in the upper middle order category are iterated through a gradient boosting algorithm which predicts the strike rate of the batsmen at different stages of the match. The average is computed and sorted in descending order.

If there are no all-rounders in top order or upper middle order, we need 2 all-rounders for balance of the team. Lower middle order batsmen are iterated through a gradient boosting algorithm which predicts the strike rate of the batsman at different stages of the match. Average is computed excluding the power play. Batsmen with the highest strike rate among them are chosen for the playing XI.

E. Bowlers *Selection*

Generally, a team plays with 2 spinners, if a spinner is available in all-rounders, we only need one spinner and the remaining 3 positions considered are fast bowlers. Economy of a bowler and bowling average are important aspects of a good bowler. Economy shows how effectively a bowler is restricting the batsmen to score runs. Average is all about taking wickets for less number of runs.

Average of the bowlers are extracted from the data set and sorted. The top 4 bowlers are iterated through KNN regression algorithm which predicts the economy of the bowler throughout the match. It is again sorted in descending order and considers the first three in the list. The same process is carried out for spin bowlers too.

Final Playing XI

As mentioned, we need input parameters such as opposition team, venue and ME. So to illustrate this model, sample inputs given are as follows:

Opposition Team = England, Venue = Lords, Minimum Experience = 15 T20I Matches

The data set goes through the process mentioned above and the final output is

V Kohli, RG Sharma, KL Rahul, S Dhawan, SS Iyer, MK Pandey, AM Rahane, JJ Bumrah, B Kumar, SN Thakur, Kuldeep Yadav

IV. CONCLUSION AND FUTURE SCOPE

Two data set files containing the data of all 132 matches played by team India cricket team are constructed. The first data set contains metadata such as Match ID (unique for each match), Team1 & Team 2 (India/opponent based on where the match has taken place), Venue, Team that won the toss, Decision of the team that won the toss, player of the match, and winner of the match. The second data set contains ball-by-ball data of all the balls bowled in those 132 matches, the fields of second data set are Match ID, Number of the innings, batsman at the striker end, batsman at the non-striker end, runs that came off the bat, Extras, and type of dismissal (if any batsman is out during that ball).

Several Machine Learning algorithms such as Linear regression, K nearest neighbor regression, Random Forest classifier, Gradient boosting algorithm, Decision trees were tested and the best algorithms were used to predict various cricket related parameters such as run rate of a team, strike rate of a batsman, economy of a bowler, number of wickets a bowler can take.

All the predicted parameters are taken into consideration to suggest a best available playing XI. The model is successful in suggesting a best playing XI, however there are a few drawbacks in the model such as not considering a new player for debut in the next to happen match while suggesting the playing XI and does not consider the consistency of the player while suggesting the playing XI.

Future Scope

The project can be extended to an advanced level, where prediction can be done ball-by-ball. In 1937, BBC became the first channel to broadcast a cricket match on the television. Since then, broadcasting technology has evolved to a new level. To give a complete watching in the stadium experience to the audience, multiple cameras have been used to capture the live. Hawk-eye technology, which is a computer vision system used to trace the trajectory of the ball. The system uses six (or sometimes seven) Full HD cameras captured at 25/30 fps, placed at different positions in the stadium.

The video feed from all the cameras is triangulated and combined to visualize the ball trajectory in a three-dimensional representation. The data of where the ball has pitched, where the ball has hit the bat, where the ball has landed after the impact will be very useful to predict the type of ball the bowler is going to bowl in the next ball, and how the batsman plays his shots against it. Such data can be algorithmically predicted, analyzed and used to characterize the batsman and bowler. Accuracy of prediction of parameters can be increased by using Hawk-eye data. If a team can guess the correct playing XI of the opponent team, a more sophisticated playing XI can be suggested based on the prediction of the match-ups such as a particular batsman VS a particular bowler, VS a type of bowling, or VS a length/line of the ball, or in a particular situation.

V. REFERENCES

[1]. Anand Vardhan, Arjun N, ShobhanaMadhavan and Deepak Gupta, “The influence of fan behaviour on the purchase Intention of Authentic Sports team Merchandise” Smart Innovation, Systems and Technologies, vol:196,pp:573-579, 2021. doi:10.1007/978-981-15-7062-9_57

[2]. Nicholas M.Watanabe,Stephen Shapiro and JorisDrayer, “Big Data and Analytics in Sports Management”, Journal of Sports Management, vol:35, pp:197-202, 2021. doi.org/10.1123/jsm.2021-0067

[3]. Ani R, Vishnu Harikumar, ArjunK.Devan and O.S.Deepa, “Victory Prediction in League of Legends using Feature Selection and Ensemble methods”, Proceedings of the International Conference on Intelligent Computing and

Control Systems- ICICCS 2020, pp:74-77. doi:10.1109/ICCS45141.2019.9065758

[4]. Sabarish B.A, Mohan S, Mamthasri D.P, Ajit R.C and Arun A.V.R, “Automating run out decisions in cricket using Image Processing”, International Journal of Applied Engineering Research, vol.10, Issue:10, pp:25493-25500, 2015

[5]. Ganesh NeelakantaIyer, BalaVignesh S, BommidiSohan, Dharmesh R and Vishal Raman, “Automated Third Umpire Decision Making in Cricket Using Machine Learning techniques”, Proceedings of the International Conference on Intelligent Computing and Control Systems-ICICCS 2020, pp: 1216-1221. doi: 10.1109/ICICCS48265.2020.9121078

[6]. Tejinder Singh, Vishal Singla and Parteek Bhatia, “Score and Winning Prediction in Cricket through Data Mining”, Proceedings of International Conference on Soft Computing Techniques and Implementations (ICSCTI), pp:60-66, October 2015. doi: 10.1109/ICSCTI.2015.7489605

[7]. I. Anik, S. Yeaser, A. G. M. I. Hossain and A. Chakrabarty, "Player’sPerformance Prediction in ODI Cricket Using Machine LearningAlgorithms," Proceedings of 4th International Conference on ElectricalEngineering and Information & Communication Technology(iCEEiCT), Dhaka, Bangladesh, 2018, pp. 500-505, doi:10.1109/CEEICT.2018.8628118.

[8]. Bunker, Rory, and Teo Sunsjak. “The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review.” SportRxiv, 2 Nov. 2020. https://www.espncricinfo.com

Cite this article as :

Madisetty Eshwar, Yasvantha Sai Atmuri, Saiteja Kalam, "Proposing Competent Team Composition for T20 Cricket Through Data Processing Techniques", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 9 Issue 5, pp. 563-570, September-October 2022.

Journal URL : https://ijsrst.com/IJSRST229574