Prediction of Final score in a Cricket Innings: A Survey

[1]Siddharth Pawar, [2]Rishabh Parakh, [3]Gaurav Patil, [4]Prof. Shashikant Mahajan
[1,2,3]UG Student, [4]Assistant Professor, Vidyalankar Institute of Technology, Mumbai, India,
[1]sidvijay83@gmail.com, [2]parakhrishabh123@gmail.com, [3]patilgaurav556@gmail.com,
[4]shashikant0320@gmail.com

Abstract- People have been passionately following sports such as cricket, football, badminton, tennis, ufc, Olympic events etc. all around the globe. A lot of people nowadays are keen to predict the final score or the winner of an ongoing sport event. Cricket is majorly played in three formats – Test, ODI and T20. This paper focuses on the Indian Premier League (IPL) which is conducted according to the T20 format of the game. It is often observed that the projected scores in an IPL T20 game are calculated using the Current Run Rate which is termed as the number of runs scored per over. But it is not an ideal overview of how an innings proceeds. Various dynamic instances within an over or two, can change the whole scenario of the innings. The aim of this survey paper is to present all the information and related works regarding the prediction of final score of an innings performed by various people. We will also look at what methodologies and machine learning algorithms they have used in order to achieve optimum accuracy for the same.

Keywords: Machine Learning, Regression, Prediction, Indian Premier League (IPL), Data Preprocessing, Data analytics.

I. INTRODUCTION

Cricket has an enormous fan base in India irrespective of whichever format it is played in. The T20 format of the game is the most celebrated one in recent years wherein both the teams get to bat as well as field for 20 overs and the team scoring more runs between the two is declared as the winner. After the introduction of Indian Premier League (IPL) in 2008, cricket has become a heavy sport accumulating investments of billion dollars as well as sponsorships from numerous popular brands in the country. The rise of IPL has fascinated a lot of people in predicting the outcomes of the IPL matches and the final score that the batting team will score in its 20 overs. One of the most luring investments in IPL months is an ever-increasing interest of the people in score prediction applications like Dream11, My11circle etc. These applications enable users to participate in certain contests upon paying the needed amount and then these applications reward the winners with cash prizes or gift coupons to those users who are successful in predicting the right outcome with respect to the contest requirements.

Usually, the calculation of the final score of the batting team in an innings is performed using the current run rate. That is a statistical method of calculating the final score. But numerous factors can change the course of the innings. Batting team scoring some quick runs by explosive batting or the bowling team taking back-to-back wickets and demoralizing the tendency of the batting team are some factors that are not considered while predicting the score using current run rate method. The runs scored by the batting team and the wickets fallen at any instantaneous moment in an innings helps us get only an initial idea about the tendency in which the innings proceeds. The reason behind selecting these factors is to build one model that can understand the dynamicity of the cricket game. Therefore, we are considering the factors which will be focusing on the dynamic changes that happen gradually in a T20 innings. The uncertainty of these factors makes data analytics followed by application of machine learning as an appropriate method for finding the solution to the above-mentioned shortcoming. In this paper, we focus on the information and other related research works available regarding the considered subject matter and thus plan on building a more accurate model that will overrule the flaw reviewed in other works.

II. LITERATURE SURVEY

Vilas Rathod and Shreyan Jain in [1] have put forward a system that displays the live updates for football and cricket matches. The authors have used cricbuzz and ESPN APIs as source of information for cricket and football respectively. They aim to make it feasible for the user to see all the live updates of cricket as well as football on the same application so that there might not be a need to install various other applications and browse different websites separately. Their proposed work is created using Java and client communication is connected through the tomcat Apache server. The system requires the user to fill the current input hence it is following a holistic approach.

 DOI : 10.35291/2454-9150.2021.0605

Prasad Thorat, Vighnesh Budhivant and Yash Shahne in [2] have introduced a system CricFirstpredictor that will help in predicting the final score of batting team in their innings of 20 overs. The authors' focus is on the Indian Premier League – as they wish to model the most dynamic format of the cricket game i.e., the T20 format. They have taken the IPL dataset from Kaggle. After cleaning the dataset, they proceeded with one hot encoding method for data preprocessing. The three machine learning techniques performed in the paper are Linear regression, Random Forest regression and Lasso regression. They developed UI using Flask framework in order to present the results. On determining the accuracy of the three approaches after calculating the Mean absolute error, Mean Squared error and the Root Mean squared error; the authors reached to the conclusion that Linear regression was the most accurate of the three – having the highest accuracy (about 76%) for the prediction.

Tejinder Singh, Vishal Singla and Prateek Bhatia in [3] aim to predict the score of both the innings of a One-day international cricket match [ODI] and thus predict the winner. The dataset has been collected from espncricinfo website and possesses all the data of matches between 2002 to 2014 for eight teams as mentioned in the related work. Further the data has been analyzed in Weka where it is divided as the training dataset and the testing dataset. The authors have implemented Linear Regression Classifier for the first innings and the Naïve Bayer's Classification for the second innings. The data of 5 overs span each from first to fifth over uptil 45^{th} to 50^{th} over has been attributed to both the playing teams for both the innings. Accuracy of Naïve Bayer's classification starts from 68 percent at the initial stage of the innings increasing to 91 percent at the end of the innings.

D Thenmozi, P Mrinalini, SM Jaisakthi et al in [4] aim to predict the winner of an ongoing IPL match at any instantaneous moment of the event. The authors collected the dataset from the website 'www.cricsheet.org' and have implemented machine learning algorithms such as 'Gaussian Naive Bayes', 'Support Vector Machine', 'K-Nearest Neighbour' and 'Random Forest regression'. After filtering the data, they implement the Recursive feature elimination method in order to sort and select the features. Generated models were based on the phases: from 2-overs to 20-overs: in an ongoing match. They analyse the results by considering few factors such as – 'the number of features selected', 'which phase of the match is inputted', 'which machine learning technique is used' and 'the team considered for prediction' – and thereafter conclude that random forest regression provided the optimum accuracy for prediction. The overall accuracy in their model is 75.5%.

A N Wickramasinghe and Roshan D. Yapa in [5] focus on predicting the man of the match as well as the outcome of that particular cricket match. The authors extracted around 150000 tweets using R studio by accessing twitter API and segregated them into negative and positive ones. 3 models proposed for the prediction were – one based on tweets, the other based on natural parameters and the third based on both tweets and natural parameters. 'Logistic Regression', 'Support Vector Machine', 'Random Forest' and 'Naïve Bayes' were the methods used for classification and the logistic regression classifier was used to find out whether there is any relationship between being popular and being elected as the man of the match. They tested all the models using 10 cross validation and later evaluated them on Accuracy, Sensitivity, Specificity and Cohen's kappa statistic. It was concluded mixed model provided the maximum accuracy of 89% while the natural parameter-based model was 83% accurate.

Aminul Islam Anik, Sakif yeaser, AG M Imam Hossain and Amitabha Chakrabarty in [6] propose a model to predict a player`s performance in an upcoming match by studying their past data. The authors focus on the players of Bangladesh National Cricket team and have collected the data from howstat and espncricinfo websites. They preprocessed the dataset and used Scikit-learn package in order to select 5 important features from it. Thereafter two methods used for feature selection were – 'Recursive Feature Elimination' and 'Feature Selection Using Univariate Selection'. They eliminated uncertain features such as 4s and 6s hit by a batsman in an upcoming match which cannot be predicted. 'Linear regression', 'Support Vector Machine (SVM) with Linear and Polynomial Kernel' and 'Processing text data with Bag of words concept' were the three techniques used for predicting the performance of any selected player. The model provided up to the mark accuracy for all the players and the highest accuracy of about 92% was achieved while predicting Tamim Iqbal`s batting performance in his upcoming match.

Siyamalan Manivannan and Mogan Kaushik in [7] have put forward two approaches – one is feature encoding and the other is Convolutional Neural Network (CNN) for predicting the outcome of the cricket match. The author sets up his proposed approach with the baseline approach. They collected the data from espncricinfo website using the python wrapper. In this paper, data augmentation is applied to training and testing dataset and normalized the final feature vector before giving it to linear SVM classifier. The authors applied clustering method during the feature encoding approach and used the player-category similarity as a weighted distance measure in order to determine the team-category relationship. While applying the CNN they connected the input nodes and the hidden nodes followed by implementing the CNN property of weight sharing. The accuracy of the model was over 70 percent using shallow CNN by training just 4000 parameters.

Manuka Madhuranga Hatharasinghe and Guhanathan Poravi in [8] aim to use computer intelligence to predict

 DOI : 10.35291/2454-9150.2021.0605

cricket match outcomes and are studying the related works for the same. This paper is a result of an ongoing research. The authors point out match stimulation, team selection and player performance analysis as the three main domains to improvise the accuracy for prediction. The two main approaches that they followed were the use of historical cricket data and the collective knowledge. In usage of historical data, classification using team and categorical data done by naïve Bayer's algorithm gave the maximum accuracy. Moreover, they discussed the team composition approach that provided the highest accuracy by using the historical data. Analyzing the existing works by relating the data used to that of the raw features, engineering features and the categorical features helped the authors conclude that the collective knowledge approach provided an accuracy of whooping 87%.

Jalaz Kumar, Rajeev Kumar and Pushpender Kumar in [9] use Decision Trees and Multilayer Perceptron Network to analyze the outcome of a cricket match. The authors have considered various pre-game and in-game attributes for the same. The data has been extracted from EspnCricinfo website by running a scraping script in a justified manner sending requests at every second. After preprocessing the data, they applied Decision tree and Multilayer Percepton Network classification techniques to it. In MLP, all vectors comprising the training samples were held in X & the target values for respective training samples were held in Y. In Decision tree technique the subsets were split and they used the impurity function for determining impurity followed by partioning the subsets until the maximum allowable depth was reached. After analyzing all the factors that strive to affect the flow of a cricket game, they compare the simulation results such as accuracy value and precision value. That helped the authors to conclude that the Multilayer Perceptron Network approach was more accurate than Decision technique displaying an accuracy of 0.576 and 0.551 respectively.

Sunil Ray in [10] provides significant information about regression analysis and states it as technique that determines a relationship between dependent and independent variable. Thereafter the author states the benefits of the regression techniques and the three metrics that drive them – 'number of dependant variables', 'shape of the regression line' and 'type of dependant variable'. Later, he explains the 7 regression techniques in detail which are Linear Regression, Logistic regression, Polynomial Regression, Stepwise Regression, Ridge Regression, Lasso Regression and ElasticNet Regression. He makes a point to focus that the regression technique to be implemented should be selected by considering the conditions of data.

III. CONCLUSION

After going through all the related works and information, one can conclude that the outcome and final score predictions for Cricket matches remains an exciting and promising area for research. T20 format is proven to be the best suitable cricketing format to model a game because many T20 leagues are played throughout the year. But achieving an optimum accuracy for predicting the exact range for the final score in a T20 innings remains a challenging task due to the fast-changing dynamics of the game. Moreover, one can conclude that it is impossible to predict the final score of an innings in a cricket match without analyzing the provided data thereafter applying machine learning techniques to it. A particular methodology or an algorithm cannot be pointed out as the best way to go about the process for achieving an accurate prediction as numerous parameters such as – which dataset is being used, how minute is the information in it, how are you preprocessing it, what attributes are being taken into consideration etc. play a vital role. Hence it is advised to perform the algorithm which provides you an optimum accuracy according to the chosen dataset. A keen point to notice – the more the number of attributes and schema in the dataset, more is the accuracy of the implementing model.

REFERENCES

[1] Vilas Rathod, Sheyann Jain. "Live Score of Sports". Issued February 2018. DOI 10.17148/IJARCCE.2018.7245.

[2] Prasad Thorat, Vighnesh Budhivant, Yash Shahne. "Cricket Score Prediction". Issued May 2021. www.IJCRT.Org

[3] Tejiender Singh, Vishal Singhla, Prateek Bhatia. "Score and winning prediction in cricket through data mining". Accessed November 2021. DOI 10.1109/ICSCTI.2015.7489605.

[4] D Thenmozi, P Mirunalini, S M Jaisakhti et al. Moneyball – "Data mining on cricket dataset". Issued October 2019. DOI 10.1109/ICCIDS.8862065.

[5] A. N Wickramasinghe, Roshan D. Yapa. "Cricket Match Outcome Prediction Using Tweets and Prediction of the Man of the Match using Social Network Analysis: Case Study Using IPL Data". Accessed November 2021. DOI 10.1109/ICTER.2018.8615563.

[6] Aminul Islam Anik, Sakif Yeaser, A.G.M Imam Hossain, Amitabha Chakrabarty. "Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms". Issued January 2019. DOI 10.1109/CEEICT.2018.8628118.

[7] Siyamalan Vanimannan, Mogan Kausik. "Convolutional Neural Network and Feature Encoding for Predicting the Outcome of Cricket Matches". Accessed November 2021. DOI 10.1109/ICIIIS47346.2019.9063316.

[8] Manuka Maduranga Hatharasinghe, Guhanathan Poravi. "Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Links". Issued March 2020. DOI 10.1109/I2CT45611.2019.9033698.

[9] Jalaz Kumar, Rajeev Kumar, Pushpender Kumar. "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks". Accessed November 2021. DOI 10.1109/ICSCCC.2018.8703301

[10] Analytics Vidhya. "7 Regression techniques you should know". Accessed November 2021. https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

 DOI : 10.35291/2454-9150.2021.0605