Final Score Prediction of an Innings using Machine Learning

[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- The most followed and celebrated events around the globe today are sport events which include Football, Cricket, UFC, Formula1, Baseball, Basketball and many more. This paper focuses on a sport called Cricket which is played primarily in 3 formats – The test format, the ODI format and the T20 format. T20 is the shortest as well as the most dynamic format of the game and many T20 leagues are played throughout the year. We are considering the teams of Indian premier League (IPL) for predicting the final score of a batting team in their quota of 20 overs. 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. We aim to create a model that can consider the dynamicity of the game and can predict the final score by providing optimum accuracy.

Keywords —Data preprocessing, Machine Learning, Data splitting, Regression, Indian Premier League (IPL), Prediction

I. INTRODUCTION

In a T20 format of cricket game, 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. Introduction of the Indian Premier League (IPL) in 2008 has made Cricket a heavy sport by attracting various popular brands to provide sponsorship in exchange to their publicity. Billions are invested while conducting and celebrating the league. Nowadays people not only are excited to follow their favorite teams on matchdays but also are keen to predict the final score and the winner of an ongoing match. This observation has led to many luring investments in fantasy applications which enable users to participate in certain contests upon paying the needed amount. The ones who are successful in predicting the right outcome as per the contest requirements are then presented with cash rewards and gift coupons.

We have gone through various other systems that predict the final scores and outcomes of a cricket match and we can observe that the projected scores of a team are calculated using the Current Run rate. That is a statistical method of calculating the final score. But numerous factors can change the scenario of an 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. Hence, we look forward to propose a model that will consider the tendency and gradual changes that happen as the innings proceeds. We have analyzed and reviewed the related works on cricket score prediction below and later discussed the methodology, implementation and results of our model.[1]

II. LITERATURE SURVEY

We analyzed and studied some related works based on the domain of our paper i.e., cricket score prediction. The inferences of all the 9 papers that have been studied are described below.

Vilas Rathod and Shreyan Jain in [2] have used Java and Apache Server to propose a system that collects information from Cricbuzz and ESPN APIs to display the live updates for football and cricket matches. Prasad Thorat, Vignesh Budhivant and Yash Shahne in [3] have introduced a system that uses Linear regression, Random Forest regression and Lasso regression to predict the final score of an innings and later display the result on a webpage developed using 'Flask'. Tejinder Singh, Vishal Singla and Prateek Bhatia in [4] predict the outcome of an ODI cricket match by analysing the data on 'Weka' and using the Linear regression classifier and Naïve Bayes Classification. D Thenmozi, P Mrinalini, SM Jaisakthi et al in [5] predict the winner of an ongoing IPL match at any instantaneous moment by selecting features using recursive feature

 DOI : 10.35291/2454-9150.2021.0617

elimination and then apply machine learning algorithms such as 'Gaussian Naive Bayes', 'Support Vector Machine', 'K-Nearest Neighbour' and 'Random Forest regression'. A N Wickramasinghe and Roshan D. Yapa in [6] implement 'Logistic Regression', 'Support Vector Machine', 'Random Forest' and 'Naïve Bayes' on dataset possessing 150000 tweets extracted from twitter API using R to predict the outcome and the man of the match for a cricket game. The authors implement sentimental analysis on the extracted tweets dividing them into positive and negative ones. Aminul Islam Aniak, Sakif yeaser, AG M Imam Hossain and Amitabha Chakrabarty in [7] propose a model that predicts the selected player`s performance in an upcoming match by feature selection process and thereafter applying classification techniques like 'Linear regression', 'Support Vector Machine (SVM) with Linear and Polynomial Kernel' and 'Processing text data with Bag of words concept'. They have considered the players of Bangladesh National Cricket team for the same. Siyamalan Manivannan and Mogan Kaushik in [8] predict the outcome of a cricket match using two methods – one is feature encoding and the other is Convolutional Neural Network (CNN). Manuka Madhuranga Hatharasinghe and Guhanathan Poravi in [9] study the related works for predicting the outcome of a cricket match using computer intelligence and conclude that the 'use of historical data' and 'collective knowledge' are two vital approaches for the same. Jalaz Kumar, Rajeev Kumar and Pushpender Kumar in [10] extract data from EspnCricinfo website and implement Decision Trees and Multilayer Perceptron Network to analyze the outcome of a cricket match. They have focused on various pre-game and in-game attributes for the same.

III. PROPOSED SYSTEM

We are proposing a system that helps the user to predict a range of final score i.e., from lower bound to upper bound by achieving optimum accuracy. Firstly, the user needs to fill the input fields provided on the webpage. The input fields require basic details of the innings that include the Batting team and the Fielding team. Then he/she needs to provide the current score of the ongoing match that includes the number of runs scored, the number of overs bowled and the number of wickets fallen - at that particular instance. Lastly the user should input the dynamic factors – the runs scored in last 5 overs and the wickets taken in the last 5 overs - that will reflect the dynamic changes and the tendency shift of the innings. Upon providing all the input fields, the system enables the user to predict the final score.

IV. METHODOLOGY

Following steps were performed while building the system.

Data Collection: The dataset has been extracted from Kaggle. It possesses data of over-to-over right from the commencement of IPL in 2008 upto the end of season 2017. It's an enormous dataset of 76015 rows and 15 columns (each representing one attribute).

Data Cleaning: From the selected dataset, some teams have not played all the seasons of the IPL and are not playing the upcoming season too. Therefore, data of inconsistent teams such as Deccan Chargers, Kochi Tuskers Kerala, Rising Pune supergiants, Gujarat Lions, Pune Warriors India etc. has not been considered for the prediction. Moreover, we have removed few unwanted columns from the dataset as they will not be useful for prediction. As we have considered attributes such as "wickets taken in last 5 overs" and "runs scored in last 5 overs", we have removed the data of first 5 overs of every inning. We have converted the 'date' column to date-time object from string format thus making it suitable for use. The attributes that will be used for prediction are described below.

Attributes	Descriptions
Bat_team	The team which is batting
Bowl_team	The team which is fielding
Runs	Runs scored by the batting team at that particular instance
Wickets	Wickets lost by the batting team at that particular instance
Overs	Overs bowled by the bowling team at that particular instance
Runs_last_5	Runs scored by the batting team in immediate previous 5 overs from the current instance
Wickets_last_5	Wickets taken by the bowling team in immediate previous 5 overs from the current instance

Table 1: Description of the attributes

Data Preprocessing: While preprocessing the data, we have implemented the 'One hot encoding' method. It is a process of converting the categorical data variables into numerical values thus making it suitable to use while implementing machine learning algorithms. Thereafter we rearranged all the columns of the dataset.

Data Splitting: Data was split into training and testing dataset where the data of the teams before year 2016 was considered for training the model and the data after 2016 was considered as the testing data.

Generating the model: The models have been generated using the Linear regression, Ridge regression and Lasso Regression. The model that provides the highest accuracy will be considered for prediction.

Presenting the Final prediction: The user inputs will be taken from the end user and the data will be passed to the model which will lead to the prediction of final score from lower bound to upper bound. score.

 DOI : 10.35291/2454-9150.2021.0617

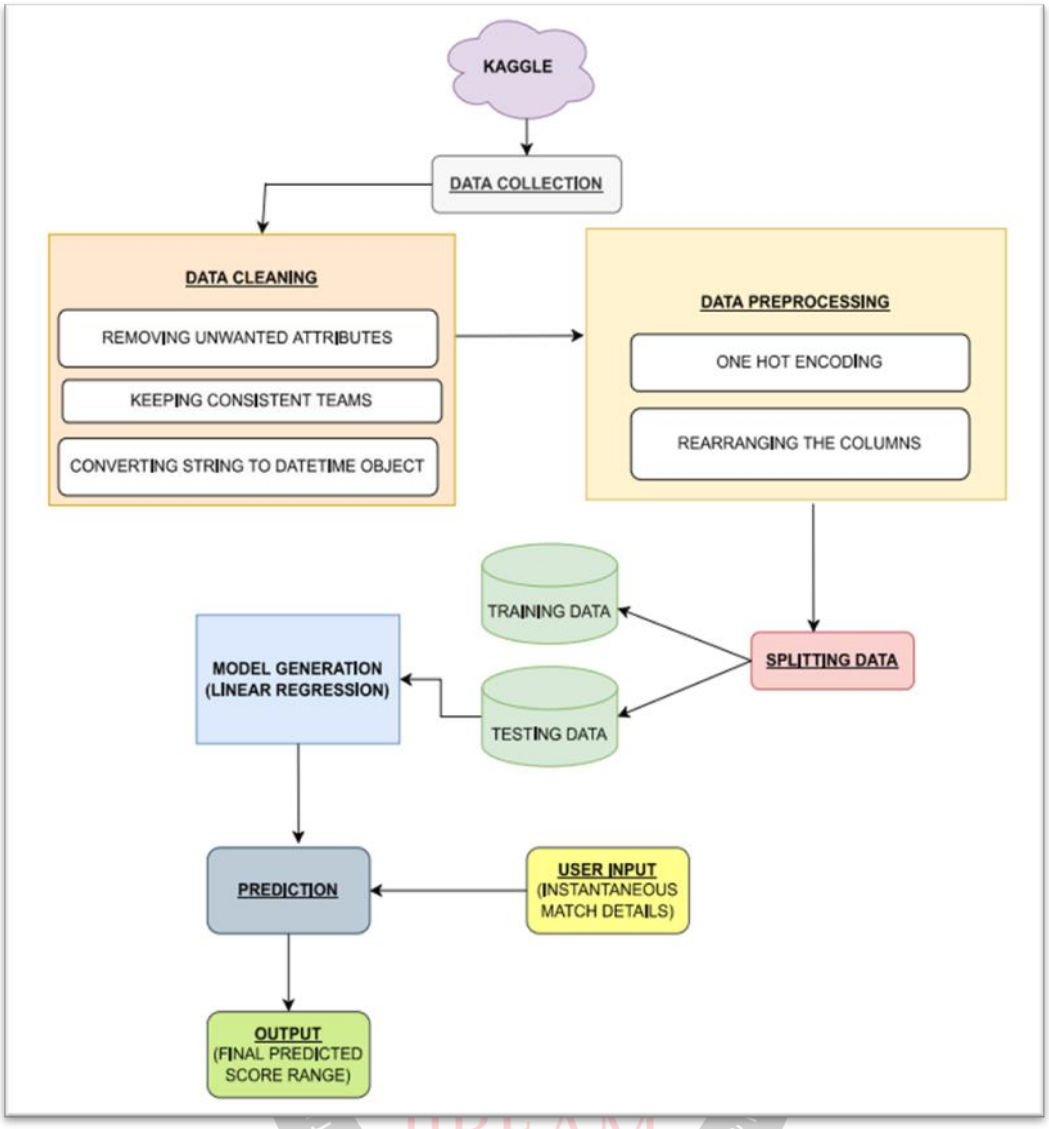

Fig 1: System Methodology Flowchart

V. IMPLEMENTATION

- **Models:**

a) Linear Regression:

Linear Regression is a Supervised Machine learning algorithm that takes the help of historical data for prediction. It helps the user to determine the relationship between the dependent and the independent variable. In the case of this project, value of the final score is the dependent variable(y) and the input factors used for prediction are the independent variables(x). The equation for Linear Regression is in the form

Y = a + bX where Y is the dependent variable and X is the explanatory i.e., the independent variable.

b) Ridge Regression:

Ridge regression is used for those models where the data suffers from multicollinearity. It is used as a shrinkage operator that reduces model complexity by shrinking the coefficient. The main assumptions of Ridge regression are linearity, independence and constant variation. [11]

c) Lasso Regression:

Lasso regression is a regularization technique which is used for a more accurate prediction. The word "LASSO" stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. Lasso regression is uses shrinkage and is suitable for models wherein methods of parameter elimination of variable selection are used. [12]

- **UI development:**

In this project, Flask framework has been used for the UI development. The main web page of the project takes the required inputs from the user in order to predict the final score. The user inputs required are Batting team, bowling team, runs scored at that current instance, overs bowled at the current instance, number of wickets at the present instance, number of runs scored in the last 5 overs and and number of wickets taken in the last 5 overs. After inputting all the fields, the user will click the "predict score" button and then the form is submitted. Model enters the scenario at

 DOI : 10.35291/2454-9150.2021.0617

the backend after the submission of the form. The inputs take the help of the historical data and are analysed through supervised machine learning techniques resulting in the prediction of final score. The routing of the pages is done based on the URLs. When the browser finds the '/' in the URL it redirects the user to the home page. After the submission of the form, the user is redirected to the '/result' URL i.e., to the result page where we can see the final result i.e., the prediction.

VI. RESULTS

We are using evaluation metrics such as MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and R squared Value for evaluating all the 3 models.

- Mean Absolute Error (MAE) is the average of difference between the actual data value and the predicted data value. It is calculated as shown below:

$$MAE = (1/n) * \Sigma|y_i - x_i|$$

where:

- **Σ:** A Greek symbol that means "sum"
- **y_i:** The observed value for the i^{th} observation
- **x_i:** The predicted value for the i^{th} observation
- **n:** The total number of observations

[13]

- Mean Squared Error is the average squared difference between the estimated values and the actual value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Where, n = Data set observations

Y_i = Observation values

$Y\hat{}_i$ = Predicted Values [14]

- Root Mean Squared Error is the root of MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$$

Where, n = Data set observations

S_i = Predicted values

O_i = Observations [15]

- R squared value is used for measuring the accuracy of the model.

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

Where,

R^2 = coefficient of determination [16]

Following are the values of evaluation metrics for the three models – Linear regression, Lasso regression and Ridge regression.

	MAE	MSE	RMSE	R Squared Value	Accuracy
Linear Regression	12.4435	258.8112	16.7781	0.7598	75.98%
Ridge Regression	12.1173	251.0317	15.8440	0.7447	74.47%
Lasso Regression	12.2141	262.3797	16.1982	0.7409	74.09%

Table 2: Values for evaluation metrics

VII. CONCLUSION AND FUTURE SCOPE

We studied and applied three regression techniques in this project. From the results, we can conclude that Linear regression is providing us the highest accuracy. Therefore, we will consider the linear regression as the prediction model for the project.

This approach can help the players as well as the management staff for making strategical decisions and selecting the appropriate team combination for the upcoming match. Cricket lovers who participate in various award-winning contests related to the prediction of the match outcomes can find this system very helpful. This methodology can not only be applied on vast datasets but also developed on bigger interfaces ahead for applications like dream11, my11circle etc. to use them and increase their brand values. Various other crucial factors like the venue, weather and pitch behavior can be considered in future systems in order to replicate as well as understand the real match scenario and thus get near-to-accurate prediction.

REFERENCES

[1] Siddharth Pawar, Rishabh Parakh, Gaurav Patil, Shashikant Mahajan. "Prediction of Final score in a Cricket Innings: A Survey". Issued December 2021. DOI: 10.35291/2454-9150.2021.0605

 DOI : 10.35291/2454-9150.2021.0617

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

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

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

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

[6] 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 December 2021. DOI 10.1109/ICTER.2018.8615563.

[7] 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.

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

[9] 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.

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

[11] https://www.mygreatlearning.com/blog/what-is-ridge-regression/

[12] https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/

[13] https://www.statology.org/mean-absolute-error-python/

[14] https://en.wikipedia.org/wiki/Mean_squared_error

[15] https://www.sciencedirect.com/topics/engineering/root-mean-squared-error/

[16] https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.htmln

 DOI : 10.35291/2454-9150.2021.0617