IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CRICKET SCORE PREDICTION

[1]Prasad Thorat, [2]Vighnesh Buddhivant, [3]Yash Sahane

[123]Student

[123]Department of Information Technology,

[123]VPPCOE & VA, Mumbai, Maharashtra, India

Abstract: Nowadays the final score of the first innings of any cricket match is predicted using CRR (Current Run Rate) method. The number of average runs scored in an over is multiplied by the total number of overs to get the final score. These kinds of systems are not useful when the T20 matches are considered because in T20 cricket the match can change its state very quickly irrespective of current run rate. The match can change within 1 or 2 overs. So, to get an accurate score prediction we should have a system that can predict the first innings score more effectively. Lots of people like watching cricket and they also like to predict the final score. This research paper focuses on an accurate prediction of cricket scores for live IPL matches considering the previous dataset available and also considers the various factors that play an important role in the score prediction.

Index Terms **- Cricket, Machine learning, Linear regression, Runs scored, Prediction, IPL, Winning probability, Random Forest Regression.**

1. INTRODUCTION

Cricket score prediction is an area where the first innings score of a cricket match is predicted using some techniques. There are various systems and prediction methods used to predict the cricket score of the ODI and the T20 cricket matches. CRR method [2] is widely used to predict the score of the cricket match. In the CRR method, the number of runs scored in an over is multiplied by the total number of overs in an innings. This method only focuses on the runs made in an over but it does not focus on the different parameters. This method can only predict the score based on the current score and not based on the various parameters. We are working to improve on the predictions by considering different parameters and to improve the accuracies of the existing systems. We are considering the T20 matches for the score prediction and we will be focusing on the live cricket score prediction.

Work proposed in [2], [5], [6] consider only the ODI matches but not T20 matches. Our aim is to predict the first innings score of the live IPL match. We have studied the systems which predict the scores of the cricket match. Most of the systems focus on the current run rate. We will be predicting the score of an IPL match by considering various factors like runs scored, overs bowled, wickets taken, etc. The reason behind selecting these features is that we need to build a model that can understand the dynamicity of the cricket game. Therefore we are considering the factors which will be focusing on the dynamic nature of the cricket game. In this study we have also plotted the graphs, which show the comparison between the predicted score and the actual runs scored.

In this research paper, there are different sections. In section 2, we have reviewed and analyzed the work done by various authors in the domain of cricket score predictions. We have explained our CricFirst Predictor (CFP) system for cricket score prediction in section 3. In section 4, we have discussed the methodology of our project i.e. the implementation in depth and results. Finally, Section 5 concludes the paper.

2. LITERATURE SURVEY

We studied papers based on the area of our research i.e. Cricket prediction. We studied 8 IEEE papers and drew some conclusions from the study. A comparison of all 8 IEEE papers has been made [1].

The work proposed in [2] deals with the score prediction of the first innings and also predicts the outcome of the match after the second innings. Linear Regression algorithm is used to predict the first innings score and outcome prediction is done by using Naive Bayes Classifier. In [3], the research aims at predicting the result of an ongoing cricket match on an over-by-over basis based on the information and data that is available from each over. The author tests the datasets on various machine learning models. It has been found that the Random Forest algorithm has the highest accuracy. The work proposed in [4] deals with the sentiments. The author predicts the outcome and man of the match by using twitter-based positive and negative sentiment analysis. Naive Bayes classifier, SVM, Random Forest algorithm, and Logistic Regression, are some of the models used for prediction. In [5], cricket squad analysis is done. This paper provides a mathematical approach to select the players. RMSE value of Multiple Random Forest Regression is greater than LR, SVR, and Decision Tree.

The work proposed in [6], deals with the player's performance prediction in ODI matches. This paper proposes the model which consists of statistical data of Bangladesh players. The main aim of this research is to predict the performance of players based on the records using SVM with linear kernel and SVM with the polynomial kernel.

In [7], CNN and Feature encoding is used to predict the outcome of cricket matches. The research predicts the outcome of a cricket match is predicted even before the match starts. The accuracy of shallow CNN is over 70%. The work proposed in [8] is the research of various papers. In this paper author analyses the work done by various authors in the cricket prediction domain. In [9], outcome prediction of ODI matches is done using decision trees and MLP networks. A comparative study is done between MLP and Decision trees and final results are given. In this study, a comparative analysis of the predictions generated by 2 different supervised classification models was performed for the same input dataset.

3. PROPOSED SYSTEM

After studying the research papers, it has been observed that the prediction is usually done without considering the dynamic nature of the cricket game. We are proposing a system CricFirst Predictor (CFP) that can consider the dynamic nature of the game. The accuracies of the models used by authors are slightly on the lower side. We are trying to improve the accuracy of the algorithm by considering various parameters.

3.1. Model Architecture:

We aim to build a model that can predict the first innings score of a live IPL match efficiently. We are looking to build a model that can consider various parameters that contribute to the score prediction.

a) Data Collection:

We will be taking the dataset from the datasets available on Kaggle. The dataset will be taken in the CSV format. The data collected from the website will be cleaned in the next step.

b) Data Cleaning:

In the data cleaning step, we want to remove unwanted columns like match id, venue, batsman name, bowler name, a score of the striker, and score of the non-striker. These columns will not be required during prediction hence we will be dropping those columns. In the IPL dataset, some teams are not playing in the IPL anymore. Teams like Deccan Chargers, Kochi Tuskers Kerala, Pune Warriors India, Gujarat Lions, Rising Pune Supergiant, etc. are not part of IPL. So, we need to eliminate those teams from the dataset and we only need to consider the consistent teams. We will be considering the data after 5 overs. The date column in the dataset is present in the string format but we want to apply some operations on the date column for that we will need to convert the string to a date-time object.

c) Data Preprocessing:

After cleaning the data, we will need our data to be preprocessed. In the data preprocessing step, we will be performing one-hot encoding. One hot encoding is explained in detail in the implementation section. We will need to rearrange the columns of our dataset in the data preprocessing step. The purpose of rearranging columns is that we need our columns to be properly arranged in some sequence.

d) Data Splitting:

After data preprocessing, we will be splitting our data in such a way that IPL matches played before 2016 will be considered for the training of the model and IPL matches played after 2016 will be considered for test data.

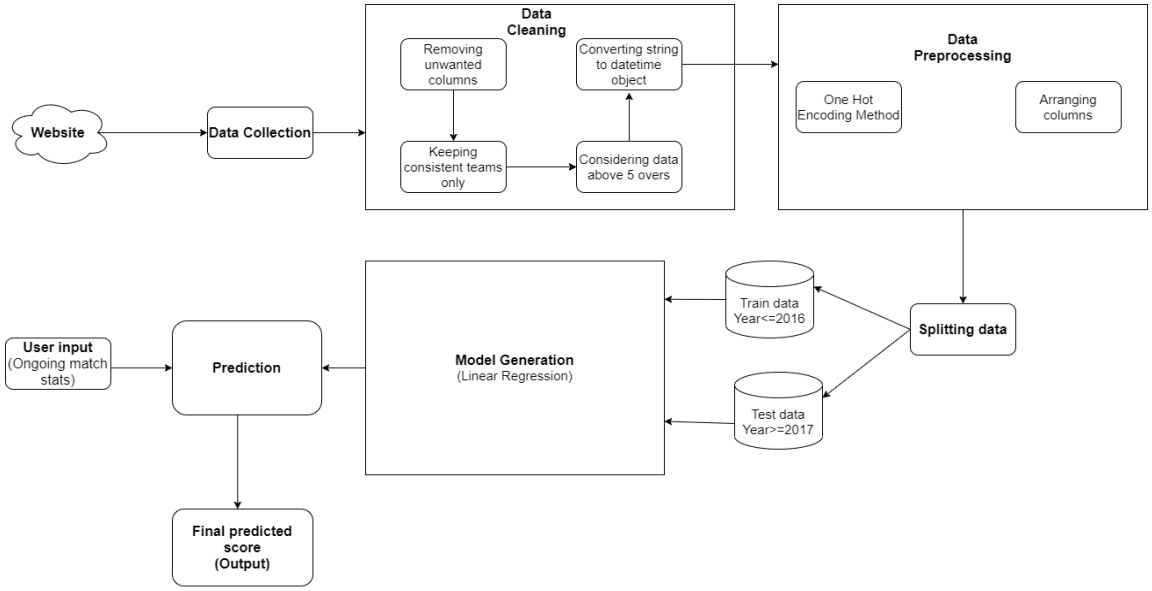
e) Model Generation:

We will be using the Linear Regression model, Random Forest Regression and Lasso Regression model for the prediction. The model with highest accuracy will be selected for the prediction. The model which we will be using for the prediction is explained in the implementation section.

f) Final Prediction:

Finally, the data will be passed through the model and then the user inputs will be taken. After getting the user inputs and matching them with the historical data we will be predicting a range of the score i.e. from lower bound to the upper bound.

Below is the model architecture for CFP system.

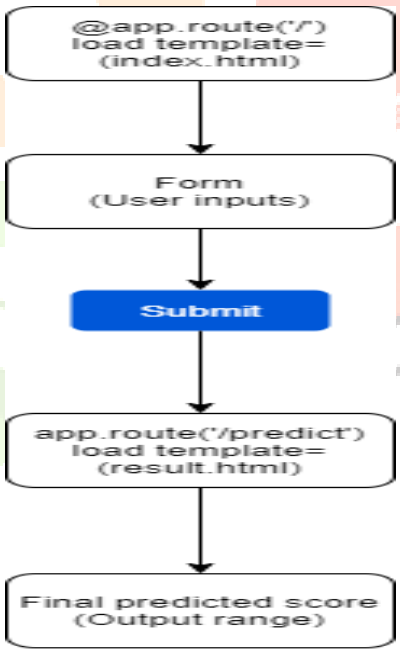


Model Architecture

fig – 1: cfp system flowchart

3.2. UI Flow:

The UI flow for our project is given in the following diagram.



UI Flow

fig – 2: ui flow of the system

We will be using the Flask framework for the development of this project. In the Flask framework, the routing of the pages is done based on the URLs. If our browser finds the '/' in the URL then we will be routing the user to the home page. On the home page, we will be having our main form. In that form, we will be taking the user inputs. The user inputs include, Name of the batting team, name of the bowling team, number of runs scored, number of overs bowled, number of wickets taken, number of runs scored in the previous 5 overs, and number of wickets taken in the previous 5 overs. Once the user hits predict score button then the form is submitted. After submission of the form, our model comes into the picture in the backend. The inputs are compared with the historical data and then a score is predicted. After the submission of the form, the user is redirected to the '/result' URL i.e. the user is redirected to the result page where the user can see the actual predicted score. On the prediction page, the user will be getting the output in the form of a range i.e. from the lower bound to the upper bound.

4. IMPLEMENTATION AND DISCUSSIONS

We performed one-hot encoding on the dataset. One-hot encoding means converting the data into a more meaningful format. Many machine learning algorithms cannot work on categorical data so we need to convert our data into numbers. In our dataset, we have the columns like batting team and bowling team. But when we give the user input to the model, our model should be able to understand the input. We have various teams in the batting and the bowling column. We do not want to give the input of the batting team and the bowling team in the string format so we perform one-hot encoding. The encoded data frame is given to the model as input.

Once the user fills the form, the data collected from the form goes to the model and the prediction takes place. We are considering Linear Regression, Random Forest Regression, and Lasso Regression as the models.

4.1. Models:

a) Linear Regression:

Linear Regression [10] is a machine learning algorithm. Linear Regression is based on supervised learning. Supervised learning means, performing predictions using historical data. There are two variables present in the Linear Regression i.e. the dependent and the independent variable. The dependent variable is the prediction variable. In our case, it is the value of the total runs (y_prediction). The independent variables(x) are used as a feature for prediction.
Linear Regression is used to find out the relation between x(input) and y(output). Following is the formula for linear regression,

$$Y = M0 + \Sigma Mi * Xi + e \text{ [10]} \qquad (1)$$

In the above equation (1) M0 is the intercept and Mi is the co-efficient that is learned during the model fitting time. Xi is the input variable and “i” denotes their number and “e” is simply an error function.

b) Random Forest Regressor:

Random Forest [11] is an ensemble technique that is used to perform regression and classification tasks. Ensemble techniques combine results of various machine learning models and give the best accurate prediction of any individual model.

c) Lasso Regression:

Lasso Regression [12] is a regularization technique. The “LASSO” stands for Least Absolute Shrinkage and Selection Operator. It is used for more accurate predictions. Lasso Regression is derived from Linear Regression.

4.2. Results and Discussions:

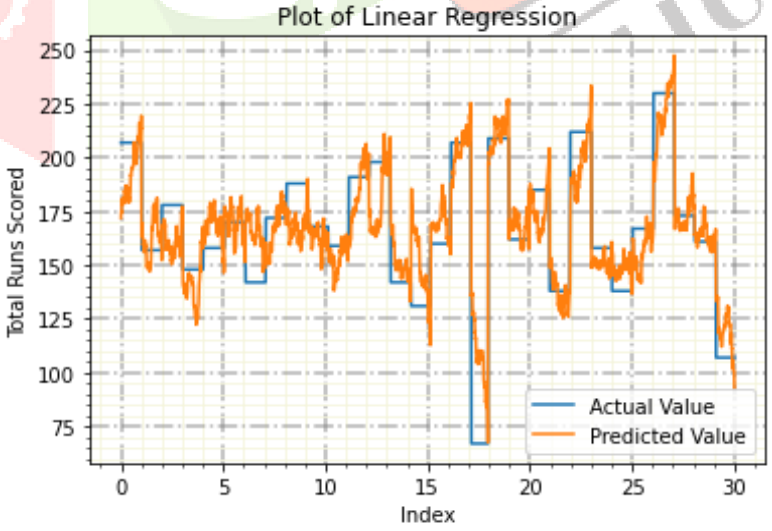


fig – 3: actual runs scored and predicted score comparison using linear regression

fig – 4: actual runs scored and predicted score comparison using random forest regression

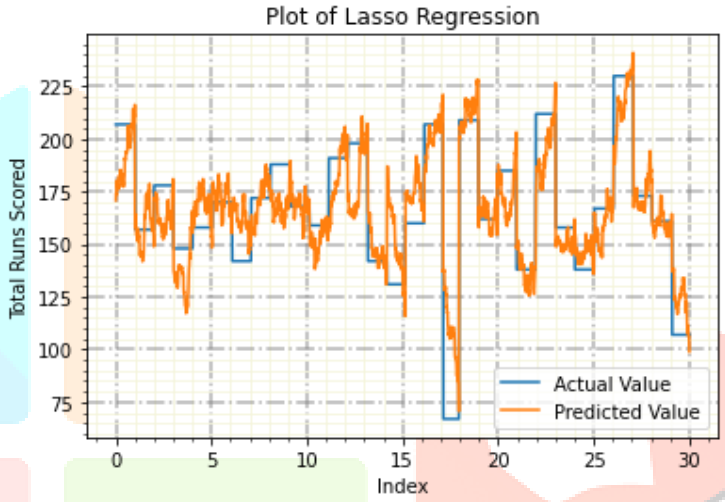


fig – 5: actual runs scored and predicted score comparison using lasso regression

The above graphs denote the actual and the predicted score analysis. The blue line denotes the actual score and the orange line denotes the predicted score. We are considering the actual and the predicted scores of 30 matches. Linear Regression and Lasso Regression graphs are almost the same. There is variation in the Random Forest Regression graph.

For evaluation of the models, we have used evaluation metrics. We are using MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and R squared value as evaluation metrics.

MAE - It is the average of the difference between the original and the predicted values.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \qquad [13] \qquad (2)$$

Where:
|xi – x| = Absolute errors.
N = Data set size.

MSE - It is the measure of the distance between the regression line and the original values and squares them to remove the negative signs.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \qquad [13] \qquad (3)$$

Where:
N = Data set size.
Y i= Observed values.

~
Yi = Predicted values.

RMSE - It is the square root of the MSE value.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad [13] \qquad (4)$$

If the MAE, MSE, RMSE values are closer to zero then we can say that the model is well fitted to the data.

R squared value - R squared value determines, how close is regression line to the original value. R squared value closer to 1, indicates that the model is well fitted to the data. In evaluation metrics, the R squared value is used to measure the accuracy of the model.

$$R^2 = 1 - \frac{RSS}{TSS} \quad [13] \qquad (5)$$

Where,
R2 = coefficient of determination.
RSS = sum of squares of residuals.
TSS = total sum of squares.

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad [13]$$

Where,
yi = ith value of the variable to be predicted.
f(xi) = predicted value of yi.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad [13]$$

Where,
yi = value in a sample.
$\bar{y}$= mean value of a sample.

We calculated the evaluation metrics for Linear Regression, Random Forest Regression, and Lasso Regression. Following are the values of evaluation metrics for the models used.

LINEAR REGRESSION

MAE value = 12.1141
MSE value = 251.6653
RMSE value = 15.8639
R squared value = 0.7516
Accuracy = R squared value * 100 = 75.16%

RANDOM FOREST REGRESSION

MAE value = 13.7799
MSE value = 330.6838
RMSE value = 18.18471
R squared value = 0.6736
Accuracy = R squared value * 100 = 67.36%

LASSO REGRESSION

MAE value = 12.2140
MSE value = 262.3797
RMSE value = 16.1981
R squared value = 0.7410
Accuracy = R squared value * 100 = 74.10%

From the above values, it is clear that the accuracy of Linear Regression is highest amongst the other models. So we will be using Linear Regression as the prediction model for this project.

5. CONCLUSION AND FUTURE SCOPE

We studied three novel approaches in this paper to predict the first innings score of a live IPL match. From the results, we can conclude that the Linear Regression algorithm has the highest accuracy of the prediction. So, we are using the Linear Regression model for the prediction purpose. Teams can use CFP to predict the final score even before the 20 overs have been bowled. So, the teams can know when to accelerate and when to play aggressively to increase the run rate while putting a target. CFP can be used to help the team management to select a team that can improve on the records. It can be used by cricket lovers for predicting the final score of a live IPL match.

In the future, we can implement a model for predicting the chasing probability. We can work on improving the accuracy of the model used in this project. Factors like venue, pitch, and the opponent team can be considered for the prediction.

ACKNOWLEDGMENT

We truly appreciate the efforts of our guide Dr. Seema Ladhe (HOD, IT, VPPCOE & VA, Mumbai, Maharashtra, India) and we are so grateful for her constant support and motivation throughout the development of this project..

REFERENCES

[1] Prasad Thorat, Vighnesh Buddhivant, Yash Sahane; Review Paper on Cricket Score Prediction; April 2021
[2] Tejinder Singh, Vishal Singla, Parteek Bhatia; - Score and Winning Prediction in Cricket through Data Mining; Oct 8-10, 2015
[3] D. Thenmozhi, P. Mirunalini, S. M. Jaisakthi, Srivatsan Vasudevan , Veeramani Kannan V, SagubarSadiq S; Moneyball - Data Mining on Cricket Dataset; 2019
[4] 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; 2018
[5] Nigel Rodrigues1, Nelson Sequeira2, Stephen Rodrigues3, Varsha Shrivastava4; Cricket Squad Analysis using multiple Random Forest Regression;2019
[6] Animal Islam Anik, Sakif yeaser, A.G.M. Emam Hussain, Amitabha Chakraborty; Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms;2018
[7] Siyamalan Manivannan, Mogan Kausik; Convolutional Neural Network and Feature Encoding for Predicting the Outcome of Cricket Matches;2019
[8] Manuka Madranga Hatharasinghe, Guhanathan Poravi Data Mining and Machine Learning in Cricket Match Outcome Prediction: Missing Link;2019
[9] Jalaz Kumar, Rajeev Kumar, Pushpender Kumar; Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks;2018
[10] https://www.geeksforgeeks.org/ml-linear-regression/
[11] https://towardsdatascience.com/machine-learning-basics-random-forest-regression-be3e1e3bb91a
[12] https://www.geeksforgeeks.org/implementation-of-lasso-regression-from-scratch-using-python/
[13]https://towardsdatascience.com/which-evaluation-metric-should-you-use-in-machine-learning-regression-problems-20cdaef258e