Cricket Score Prediction using Machine Learning

D. Akhil Kumar[a], Isha Singla[a], Gaurav Deswal[a]

[a]Department of computer Science and Engineering Chandigarh College of Engineering and Technology, India

ARTICLE INFO

Keywords:
Current score prediction
Machine learning
Deep learning
Multiple linear regression
Random forest regression
Artificial neural networks

ABSTRACT

Currently, in T20 cricket matches first innings score is predicted based on current run rate during the ongoing match. It does not take into consideration other factors like cumulative overs, cumulative runs, wickets left, etc. Thus, in this paper a method has been proposed to predict the score of the first innings using machine learning. Before proposing this method, a comparative analysis has been done using multiple linear regression, random forest regression and artificial neural networks. It has been found in the results that error in random forest regression is less than current rate method in estimating the final score.

1. INTRODUCTION

Cricket is a one of the most popular games played in many countries of the world. Its origin can be traced back in south –east England and it has developed over the centuries to become what is now one of the most loved sports with billions of fans all over the world. British brought this game to India, with the first game played in the Indian subcontinent is reported to be played in 1721. It is played by officially by major countries all over the world. Some of the countries are India, Australia, South Africa, Bangladesh and Sri Lanka.

This game is played on a field whose dimensions can vary from ground to ground but the rectangular pitch is of 22 yards (20.12 m) in length and 10 feet (3.05 m) in width always. It is played between two teams with 11 players each with a twelfth man in case one of them is injured.

The general composition of a cricket team consists of four batsmen, three all-rounders, one wicket keeper and three bowlers. It is not necessary to have such a composition but mostly it is tried to be kept like this to win the game. A batsman is a cricket player whose main role in the team is to score runs. A batsman scores runs by hitting the cricket ball with his bat and then running between the wickets whereas, a bowler is the player whose main role is to bowl at the batsman with the cricket ball. Bowlers are generally either fast bowlers or spinners. The former is the one who runs during the delivery and the latter is the one who imparts rotation or spin the ball. An all-rounder who can both bat and ball. The player who stands behind the wickets when the bowler bowl is the wicket keeper.

Cricket is played in three formats internationally. In test cricket, the two teams play a four-inning match. The match may last for five days. The first official international match was played in 1877 between England and Australia at Melbourne Cricket Ground. In one-day cricket, the two teams play a two-inning match. The match lasts for one day. The first official international one-day match was played in 1871 between England and Australia at Melbourne Cricket Ground at that time it was played as 40-40 overs match and an 8 ball over. Presently it is played as 50-50 overs format. In T-20 format**,** the two teams play a two inning match with 20 overs in each inning. The match lasts for four to five hours. It was officially introduced by the England and Wales Cricket Board in 2003.

The Indian Premier League (IPL) is a professional twenty-twenty cricket league in India played during April and May of every year by teams representing 8 Indian cities and states. The league was established by the Board of Control for Cricket in India (BCCI) in year 2008. IPL has an exclusive window in ICC future tours programme. Chennai Super Kings, Delhi Capitals, Kings XI Punjab, Kolkata Knight Riders, Mumbai Indians, Rajasthan Royals, Royal Challengers Bangalore, Royal Challengers Bangalore are the teams of IPL.

Presently in an IPL cricket match, the projected scores are displayed on the scoreboard during the first innings. During a match, predictions are made mostly on the basis of the current run rate of the team batting. Run rate

is defined as the amount of runs scored per number of overs bowled. But there are possibilities to improve this method by including other factors affecting the final score of the team such as cumulative overs, cumulative sum, cumulative strike rate, wickets left .

1.1 Motivation

Cricket is becoming one of the most famous games all over the world. It makes the use of latest technology and with changing times more and more technology updates are being introduced to improve the game. The present method for the prediction of score does not take into account many factors such as cumulative overs, cumulative sum, cumulative strike rate and wickets left.

In case, if a team makes 100 runs in 10 overs then the current method predicts 200 runs by the end of the first innings (T20 format), but if the current batsman is the last player and gets out in the next ball, then this method gives a difference of 100 runs. The difference is so because present method does not take into account the number of wickets left, etc. Hence, in this unpredictable game the score prediction should not be done using only one parameter. This paper paves a way for doing so by making use of machine learning methods for prediction of score in IPL or any T20 format as it takes into account other factors like cumulative runs, cumulative overs, cumulative strike rate and wickets left.

1.2 Contribution

Through this work, a new benchmark in the domain of cricket score prediction has been set as this work uses runs, wickets, overs and performance for the collection of data and uses machine learning models (regressions and neural networks) for the training, which is also new to this field with such a dataset. The aforementioned factors are helpful in predicting the final score of the team playing in the first innings of T20 cricket match. Another contribution of this work is the dataset. The dataset has been generated by manual data entry from various cricket score websites. Since there is no pre-existing dataset of cricket score prediction within over ranges (0-4, 5-8, 9-12, 13-16, 17-20) over the internet, web scrapers were not used in data collection. Thus, this can also help in future researches in this field.

2. RELATED WORK

Few people have worked in this field of predicting the scores or the outcome of an IPL match or in any T-20 format. Most of them have worked in the field of ODI. One of them is the work done by Scott Brooker and Seamus Hogan at University of Canterbury [5] as part of the PhD research project. It has drawn an estimate on how well the average batting teams do against the average bowling team under the listed conditions along with the current state of the game. In the first-innings, it makes an estimate of the additional runs that can be scored with the given number of balls and the number of wickets remaining. In the second innings, the method estimates the winning probability with the given number of balls and the number of wickets remaining along with the runs scored in the current situation and the target given to them by the other team. The estimation has been done with the help of dynamic programming.

Another work in the field of ODI is done by T.Singh et al from Thappar University, Patiala in their work in which the score prediction of the first innings is done on the basis of number of wickets fallen, venue of the match and batting team. The method also predicts the outcome of the match in the second innings considering the same attributes as of the former method along with the target given to the batting team. The prediction has been done based on linear regression and Naïve Bayes Classifier for first innings and second innings respectively.

Likewise, the analysis of the Indian cricket team's ODI matches data and application of association rules on the attributes namely home or away game, toss, batting first or second and the match result by Raj and Padma. Swartz et al.[3] made the use of Markov Chain Monte Carlo methods for the simulation of ball by ball outcome of a match with the usage of a Bayesian Latent variable model. Kaluarachchi and Varde[4] did the implementation of both Naive Bayes classier and association rules and did the analysis of factors that contribute to a win. However, this paper does not talk about the estimation of final score of the innings.

There have been other research on prediction of results of match in other games like football, baseball, basketball, etc. For interest for basketball, an advanced Scout System for the identification of many trends in this game has been proposed by Bhandari et al[5]. Also in the field of football, prediction of the outcome of 2006 World Cup FIFA matches has been done by Luckner et al. Gartheepan et al[8]. also made a data driven model that helps to t. mfind out when to `pull a starting pitcher'. A model of selecting the players' combination that is most appropriate for winning the games by Schultz in his paper.

As explained above, the prediction of the final score at the end of the match of the first innings is normally done based on current run rate. In this paper, a method has been proposed to predict final score of the team playing in the first innings of T20 cricket match on the basis of factors such as cumulative overs, cumulative sum, cumulative strike rate and wickets left at the end of

any over in the match. The predictions are made on the basis of two methods linear regression and random forest regression. Apart from using run rate as a factor for prediction, the number of wickets left, team performance and overs already played by the team and cumulative runs till the end of the current over are also taken into consideration.

3. METHODS

3.1 Regression

Regression is a method in machine learning where a target value is to be modeled on the basis of some independent variables. The main usage of this method is to find a cause effect relationship between the variables and also used for forecasting. The comparison between the techniques is made on the basis on the number of independent variables and type of relationship between the dependent and independent variables.

Linear Regression is a model where there is only a single independent variable (x) and a single dependent variable (y) in a two dimensional space and tries to find the best line that fits through the points. It is done as to predict the result as accurately as possible. The word simple refers to the fact that only one variable is there as the predictor[12].

The relation is built on a simple linear equation:

$$y = c_1 + c_2 * x$$

Here:
y is the dependent variable
x is the independent variable
$\mathbf{c_1}$ and $\mathbf{c_2}$ are the constants

The main objective of the linear regression algorithm is to find the best values for c_1 and c_2 as shown in Fig. 1.

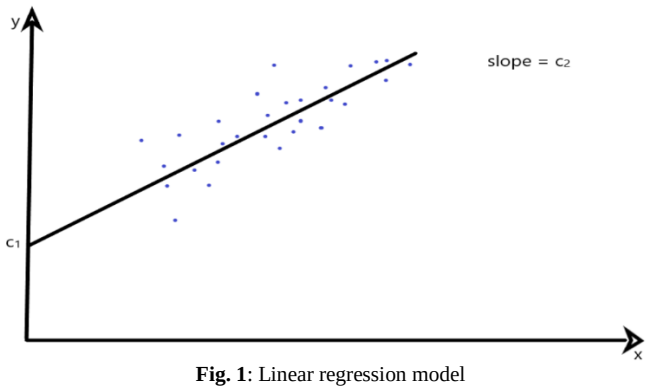

Fig. 1: Linear regression model

Multiple Linear Regression, also known simply as multiple regression, is a technique that uses several independent variables to predict the outcome of a dependent variable. The goal of this multiple linear regression model is to find out a linear relation between these variables. It is basically the extension of simple linear regression model.

The model for Multiple Linear Regression:

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + ... + \beta_p x_{i_p} + \varepsilon$$

For i = n observations, where:
$\mathbf{y_i}$ is the dependent variable.
$\mathbf{x_i}$ are the explanatory variables.
$\boldsymbol{\beta_0}$ is the y-intercept (constant term).
$\boldsymbol{\beta_p}$ are the slope coefficients for each explanatory variable.
$\boldsymbol{\varepsilon}$ is the model's error term (also known as the residuals).

A simple linear regression is a function that allows us to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

The multiple regression model is based on the following assumptions: there is a linear relationship between the dependent variables and the independent variables; the independent variables are not too highly correlated with each other; $\mathbf{y_i}$ observations are selected independently and randomly from the population; Residuals should be normally distributed with a mean of **0** and variance **σ**. The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the multiple linear regression model even though the predictors may not be related to the outcome variable. R^2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of a multiple regression model, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

3.2 Random forest regression

The random forest model is one of the most effective models in machine learning for prediction of data on the basis of some given data. This has made it one of the most famous models in practice all over the world [11]. It is basically an additive model which makes the predictions by the combination of decisions from a sequence of base models.[11]This can be represented as follows:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots$$

$g(x)$: final model

$f_i(x)$: i^{th} simple base model

Every base model is also a simple decision tree. This is one of the most basic examples of model ensemble, which involves the technique of combining the models to obtain a better predictive performance. Random forest algorithm is a supervised classification algorithm. This algorithm creates the forest with a large number of trees. The more number of trees, the more robust the model. In addition, more the number of trees ensure higher accuracy of results

The first step in random forest regression is to have multiple decision trees in a model. Next, if we want to classify a new object based on new attributes, tree is used which gives a classification and we say that tree votes for that class. The forest chooses the classifications having the most votes of all the other trees in the forest and takes the average difference from the output of different trees. In general, random forest builds multiple trees and combines them together to get a more accurate result. In the process of creating random trees, the trees are split into different nodes or subsets. The next step is to search for the best outcome from random subsets. The result is a better model for the given test[11].

Suppose we formed a thousand random trees to form the random forest to detect a 'hand'. Each random forest will predict the different outcomes or the classes for the same test features. A small subset of the forest will look at the random set of features, for example, hand or fingers. For example, if random forest regression is used for the prediction of whether it is a thumb or a finger then it makes the prediction using different random trees. It then checks the vote from each decision tree. If votes of the finger are higher, then the final random forest will return the finger as a predicted target. This type of voting is called majority voting. The same applies to the rest of the fingers of the hand, if the algorithm predicts the rest of the fingers to be fingers of a hand, then the high-level decision tree can vote that an image is a 'hand'. This is why the random forest is also known as Ensemble machine learning algorithm.

3.3 Artificial neural networks

The idea of artificial neural networks (ANN) is based on the belief that the working of the human brain can be imitated using silicon and wires as living neurons. ANN is composed of multiple nodes and is used to imitate the biological neurons of human brain.[14] The neurons are connected with the help of links so that they can interact with each other; Fig. 2 depicts the flow of ANN.

An ANN consists of the input layer, hidden layer and the output layer. The input layer is basically the data that is provided to the artificial neural network. The Input layer communicates with the external environment that presents a pattern to the neural network. The hidden layer is the collection of neurons which has activation function applied on it and it is an intermediate layer found between the input layer and the output layer. Its job is to process the inputs obtained by its previous layer. So it is the layer which is responsible extracting the required features from the input data.

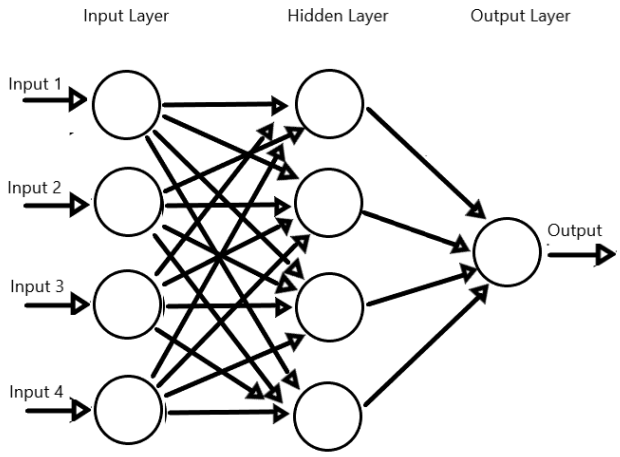

Fig. 2: ANN model

The output layer is where the computations done in the previous layers are used to determine the results. The output layer of the neural network collects and transmits the information accordingly in way it has been designed to give. The pattern presented by the output layer can be directly traced back to the input layer. The number of neurons in output layer should be directly related to the type of work that the neural network was performing [13].

Fig. 3 depicts the flow of activation function. [13] The activation function acts as a gateway between the input, which feed the current neuron, and its output, which in turn acts as an input to another neuron. There are many activation functions available which can be very simple that is binary or a complex one.

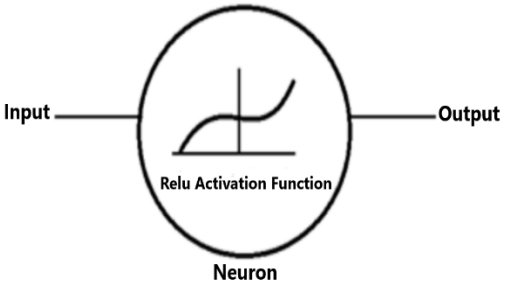

Fig. 3: Activation function

4 DATA COLLECTION AND PREPARATION

4.1 Data acquisition

Data has been collected from the website http://www.espncricinfo.com . On this website, over wise match data is available. The collected data consists of the information of the first innings of the IPL matches from the year 2008 to year 2018 of various teams.

Table 1 consists of the following attributes and their descriptions.

ATTRIBUTE	DESCRIPTION
Over range	20 overs in a T-20 format are divided into five groups as A: 1-4 B: 5-8 C: 9-12 D: 13-16 E: 17-20
cumulative overs	Total overs played after over range. (i.e. upper limit of over range)
Total runs in over range	It describes the runs scored in each over range.
Cumulative sum	It describes the total runs scored after 4 over gaps.
S/R	S/R: Strike Rate This is calculated as: ((total runs) * 100) / (6*(total overs))
wickets left	The wickets left in the match after every 4 overs.
Predicted score	The projected score according to a formula listed later.
Total runs in match	The actual total runs that the team has scored at the end of the batting.

Table 1: Attributes

First the data of the 20 overs is divided into 5 groups {A,B,C,D,E} each of which consists of 4 overs and the wickets left after the end of group of 4 overs is also entered. The total runs in each over range group is calculated and entered into the tuples. The cumulative sum of each over is calculated and entered into the tuples. Next, S/R is calculated using the formula:
S/R = ((total runs) * 100) / (6*(total overs)). The average predicted score at the end of match according o the over range is also calculated. Using the above dataset further the process of division into training and testing data is done which is explained in the next section.

4.2 Training and testing data

The dataset that has been prepared is further divided into two parts: training data and testing data. The training data consists of the data of first innings of IPL matches from year 2008 to year 2017. The testing data consists of the data of first innings of IPL matches of the year 2018. This data is further divide into two parts for the application of the regression techniques:

INDEPENDENT VARIABLES	DEPENDENT VARIABLE
cumulative_overs, cumulative_sum, cumulative_S/R, wickets_left	predicted_score

Table 2: Independent and Dependent variables

We can describe the usage of data in the following steps. First the data is trained using the training dataset.

The testing is done on the testing dataset. The case processing summary is shown as follows.

Case Processing Summary			
		N	Percent
Sample	Training	231	70%
	Testing	109	30%
Valid		340	100.0%
Excluded		0	
Total		340	

Table 3: case processing summary by ANN

The case summary in Table 3 gives the idea about the training and testing samples and also their validity. This has been used for all the samples.

4.3 Methodology

In this section, we present the methodology used for the proposed work. The objectives of the proposed work is to find efficient and accurate models for score prediction using machine learning and deep learning models and to evaluate the proposed technique with respect to the existing method. In order to carry out the proposed work, we have chosen the regressions and neural networks on score prediction since it is state-of-the-art technology that can be used to solve most of the problems be it classification in almost every domain. Thus, our plan was to collect data containing cricket score after over ranges, wickets left and overs left. After data acquisition, the data was preprocessed to remove noisy data to select only the important attributes, which results in good prediction. Then, regressions and neural network models were trained using the training dataset. Finally, prediction using regressions and neural network models was done and the results were compared. Fig. 4 depicts the workflow of the proposed work.

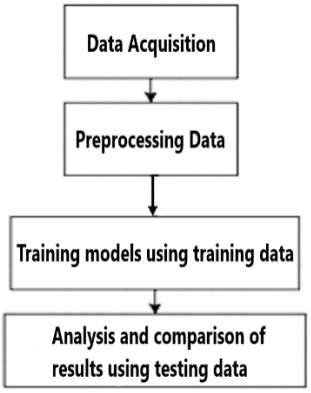

Fig. 4: Workflow of proposed method.

5 RESULTS AND DISCUSSION:

To find out the results, which have accuracy and can validate our objective, we make the use of three techniques:

5.1 Multiple Linear Regression
5.2 Random Forest Classification
5.3 Artificial Neural Networks

Experiment 1: Using multiple linear regression

Multiple regression as explained above involves the use of many independent variables and a dependent variable, which is to be predicted. The independent and dependent variables that are listed in Table 2. On application of the multiple regression model, the outcome can be summarized in the form of an equation shown below:

predicted_score = -7.2*cumulative_overs1 + 0.94*cummulative_sum + 0.04*cumulative_S/R + 3.38*wickets_left + 134.22

There are some factors, which have more effect on the total predicted score as compared to others. It can be seen that the cumulative strike rate, which takes priority in the current method, is the one that has the smallest coefficient multiplied to it in the proposed method. Along with it we can also infer that the cumulative overs, cumulative sum and wickets left are other three important factors which gain importance in this method.

The proposed model i.e. multiple linear regression is better than the current model used for run prediction can be shown using the total score prediction graphs, Error graphs and Root Mean Square Error.

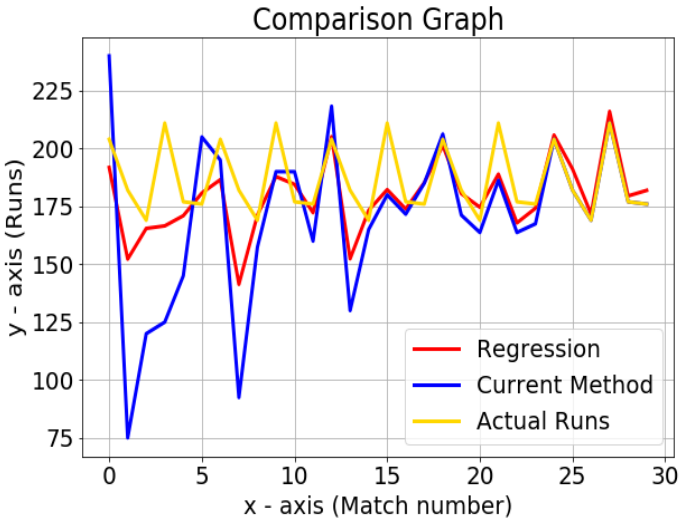

Fig 5: predicted score comparison

The Fig. 5 gives the validation that the proposed method using multiple regression is better than earlier used as the prediction by the proposed method is much more closer than that of the current method as can be seen in graph. The red line is much closer to the actual score rather than the current method used as represented by the blue line. During the initial overs (0-7), our proposed method is better than the current method. The blue line is overlapped by the yellow line between the range 22-30 matches on the X-axis.

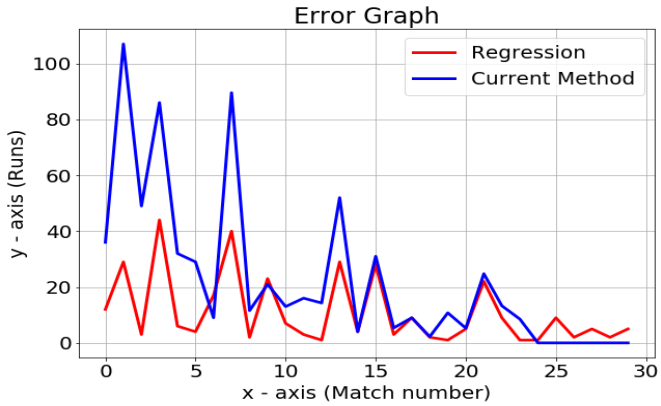

Fig 6: error comparison in predicted score

Proposed method:

Error = |Regression score– actual run scored|

Current method:

Error = |current method score – actual run scored|

	Proposed method	**Current method**
Absolute error	3.888753006517012	8.446112600536201
RMSE	16.57548644494688	35.83013405873144

Table 4: The absolute and Root mean square error of multiple linear regression model

As can be seen from the Fig 6, it can be inferred that the error in the proposed method (approx 3.9) is less than half of the error in the current method(approx 8.5)

As mentioned in (1) and (2) The root mean square error of the proposed method (around 16.5) is less than half of the root mean square error of the current method(around 35.8).

Experiment 2: Random forest regression

The experiment was performed using random forest regression on the data. Random forest regression is an ensemble learning method capable of performing regression and classification tasks and uses multiple decision trees for this purpose. The independent and dependent variables are mentioned in Table 2. On application of the random forest regression model, the graphs and the root mean square error(RSME).

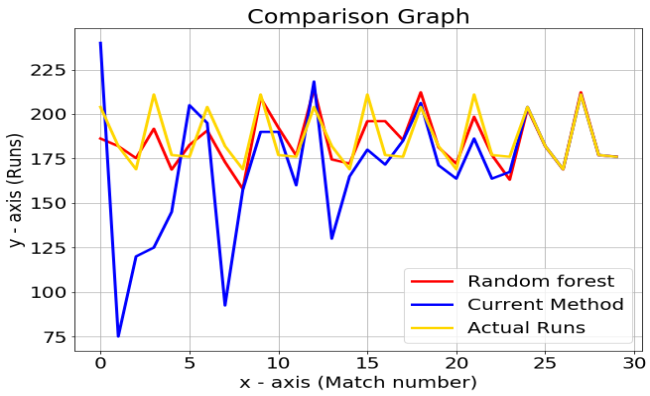

Fig 7: predicted score comparison

The Fig 7 compares the proposed method using the random forest method and the current method being used. It has been found that the proposed method perfomed better than the current method. The red line is much closer to the actual score rather than the current method used as represented by the blue line. During the initial overs (0-7), the proposed method is better than the current method. Also the closeness to the actual score keeps on increasing as we move further on the X-axis. The blue line is overlapped by the yellow line between the range 22-30 matches on the X-axis.

As can be seen from the Fig 8, it can be inferred that the error in the random forest regression method (approx. 0.9) is almost negligible. On the other hand, error in the current method is 8.5 approximately.

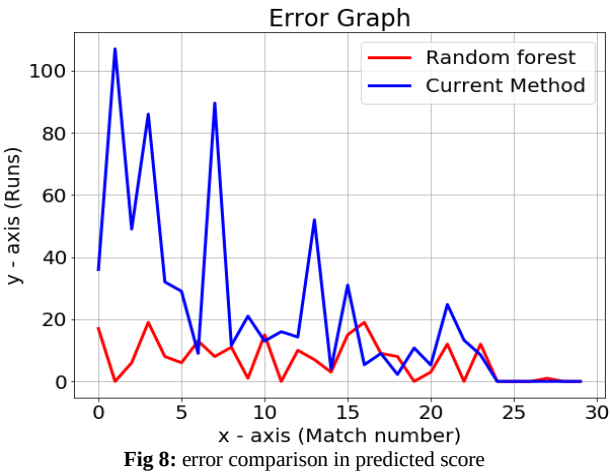

Fig 8: error comparison in predicted score

Proposed method:

Error = |Random forest score– actual run scored|

Current method:

Error = |current method score – actual run scored|

	Proposed method	**Current method**
Absolute error	0.8360781309842897	8.446112600536201
RMSE	9.488426256025974	35.83013405873144

Table 5: The absolute and Root mean square error of random forest regression model

As mentioned in Table 5 The Root Mean Square Error of the random forest regression model (around 9.5) is one-fourth of the Root Mean Square Error of the current method (around 35.8).

Experiment 3: Artificial Neural Networks

The third experiment has been performed by using ANN on the data. Artificial neural networks make the independent and dependent variables that are listed as follows in Table 2.

Table 3 gives the number and percentage of the training and testing data. The Training data comprises of 70% and testing data comprises of 30% of the total data collected. There is no exclusion of any data entry.

The artificial neural network model shown in figure 4 consists of mainly 3 layers – Input, hidden and output layer. The model as shown below also consists of 3 parts – *The input layer* in this model consists of the dependent variables – cumulative overs, cumulative strike rate, cumulative runs and wickets left. *The hidden layers* which forms the most important part of the ANN model. There are two hidden layers. The first hidden layer consists of 3 units and the second hidden layer consists of 2 units. It uses the hyperbolic tangent as the activation function. *The output layer:* The output layer consists of the dependent variable, which is the total run in the match or as described above, the predicted score, which is the major objective of this model.

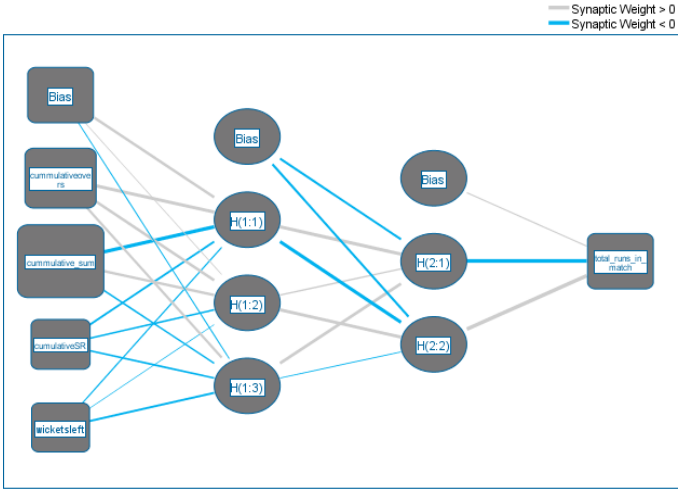

Fig. 9 Network layers

Network Information			
Input Layer	Covariates	1	Cummulativeovers
		2	cummulative_sum
		3	cumulativeSR
		4	Wicketsleft
	Number of Units		4
	Rescaling Method for Covariates		Normalized
Hidden Layer(s)	Number of Hidden Layers		2
	Number of Units in Hidden Layer 1		3
	Number of Units in Hidden Layer 2		2
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	total_runs_in_match
	Number of Units		1
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares

Table 6: Network Information

The Table 6 summarizes the proposed ANN model. It gives the information about the covariates, which are the dependent variables and the input layer. Apart from this, this table also gives information about the number of units in each of the hidden layers and the variables, which form the output layer.

Independent Variable Importance, the predictions done help in inferring some important facts as to which of the independent variables have more importance in determining the results. This can be explained using the Table 7, which is given below.

Independent Variable Importance		
	Importance (Range: 0-1)	Normalized Importance (Range: 0-100)
cumulative_overs	.276	62.8%
cumulative_sum	.440	100.0%
cumulative_SR	.147	33.4%
Wickets_left	.136	30.9%

Table 7: Independent variable importance

The cumulative sum of the runs at the end of each over holds the highest importance with the wickets left holding the least importance. The cumulative overs and cumulative strike rate have more effect as compared

to wickets left. Among these cumulative overs hold higher importance.

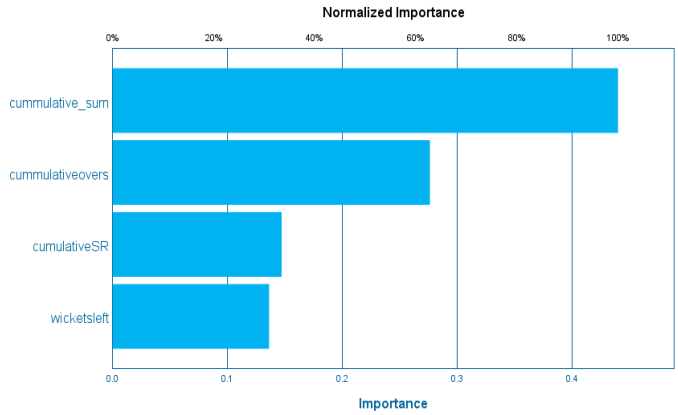

Fig. 10: Independent variable chart

The independent variable chart (Fig. 10) easily summarizes the results explained above in the following manner:

Cumulative sum > cumulative overs > cumulative strike rate > wickets left

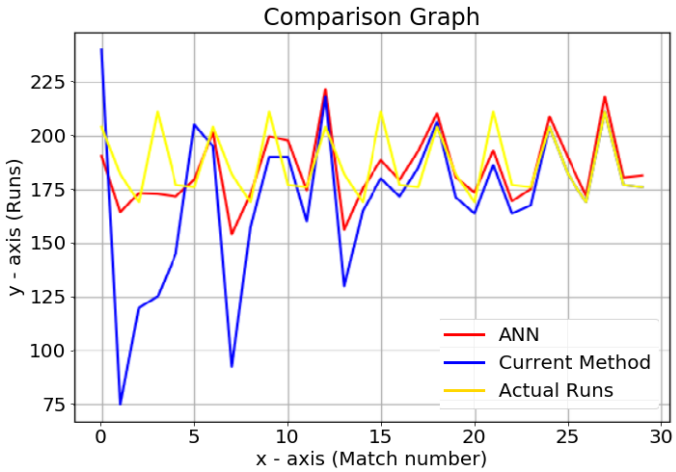

Fig 10: predicted score comparison

The Fig 11 gives the validation that the ANN method is better than regression but not with random forest. This method is much closer than that of the current method as can be seen in graph. The red line is much closer to the actual score rather than the current method used as represented by the blue line. During the initial overs (0-7), our proposed method is better than the current method. Also the closeness to the actual score keeps on increasing as we move further on the X-axis. The blue line is overlapped by the yellow line between the range 22-30 matches on the X-axis.

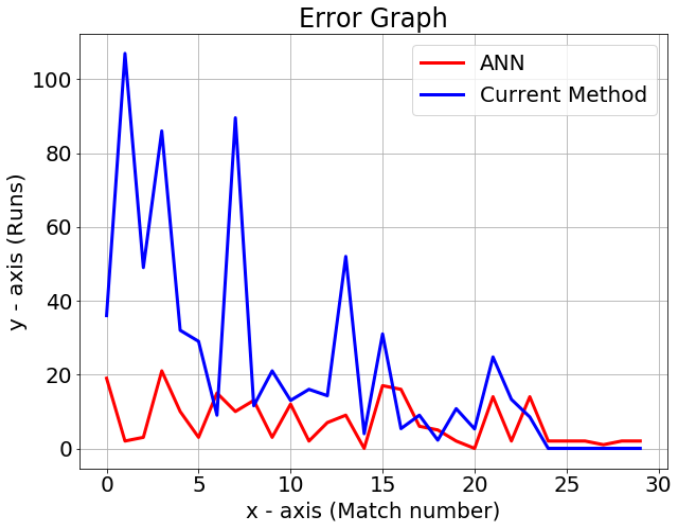

Fig 11: error comparison in predicted score

Proposed method:

Error = |ANN score– actual run scored|

Current method:

Error = |current method score – actual run scored|

	Proposed method	**Current method**
Absolute error	0.8360781309842897	8.446112600536201
RMSE	13.791254208564252	35.83013405873144

Table 8: The absolute and Root mean square error of ANN model

As can be seen from the Fig 12, it can be inferred that the error in the ANN (approx. 0.9) is almost negligible. On the other hand, error in the current method is 8.5 approximately.

It can be seen from the table that the root mean square error of the artificial neural networks model (around 14) is half of the root mean square error of the current method (around 35.8).

Model Summary

Model Summary		
Training	Sum of Squares Error	49.748
	Relative Error	.433
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.07
Testing	Sum of Squares Error	30.822
	Relative Error	.510

Table 9: independent variable importance

Table 11 summarizes the model giving all the information about training and testing errors.

6. COMPARATIVE ANALYSIS

Comparative analysis between multiple regression model, random regression model and artificial neural network has been done and it can be inferred that random forest regression gives the best results.

Random Forest is much more efficient in the discovery of more complex dependencies at the cost of more time for fitting. If there is a presence of a linear relationship between the dependent and independent variables then there is a possibility that the results of multiple linear regression and random forest regression are similar. But, if the dependency is something different from linear, there is no possibility for an algorithm to come up with a linear equation.

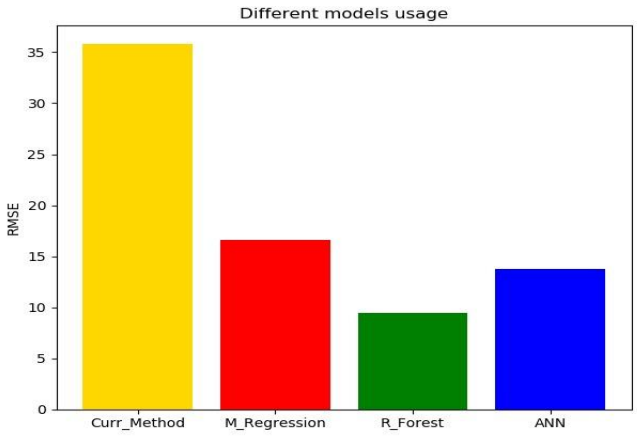

Fig 12: RMSE of models

Also, since random forest regression makes the use of various smaller subsets therefore it is much capable of giving better results. It can be inferred from the Fig 12 that the Root Mean Square Error of Random Forest Regression is less than that of all the other three. Thus, random forest regression is best for the prediction of total score.

5.1.1 Case Study:

A case study is performed on a match number 12 played on March 31, 2019 between Chennai Super kings and Rajasthan Royals in IPL 2019.

Using random forest model we predicted the score of Chennai super kings in the first innings and the results are shown in table 9.

Table 9: predicted score by random forest and current method

Overs	Score predicted using Current method	Score predicted using Random forest	Final score scored at the end of the match
After 4 overs	110	157	175
After 8 overs	95	142	175
After 12 overs	130	152	175
After 16 overs	135	161	175

Fig 13 shows the comparative analysis between the existing method, the actual runs made in those matches and the random forest regression. It can be inferred from the graph that the red line in closer to the yellow line as compared to the blue line. Table 9 and Fig 14 clearly depicts that method using random forest model's prediction is very close to actual final score scored by the team in the first innings.

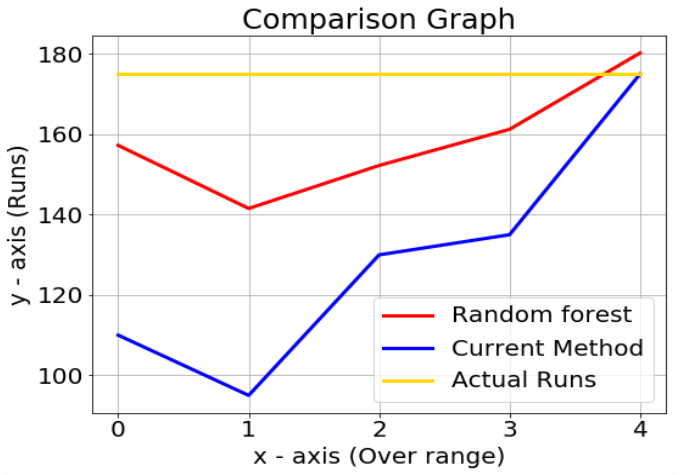

Fig 13: predicted score comparison

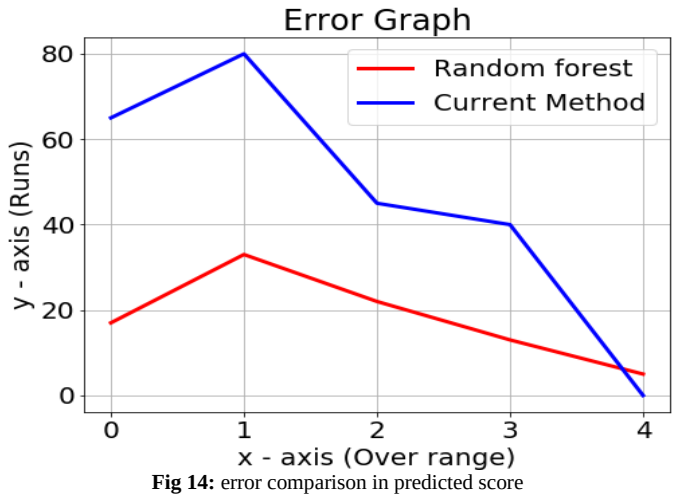

Fig 14: error comparison in predicted score

The red line is much closer to the actual score rather than the current method used as represented by the blue line. Also the closeness to the actual score keeps on

increasing as we move further on the X-axis. From Fig 14, it can be inferred that the error in the Random forest is less than to current method.

ROOT MEAN SQUARE ERROR (RMSE):
RMSE: 17.46268656716418 (Random forest)
RMSE: 52.68407960199005 (Current Method)

7. CONCLUSION

The main purpose of this paper is to make a model for predicting the final score of the first innings. Three models which are multiple linear regression, random forest regression and artificial neural networks have been tested to find the best one out of them. After the comparative analysis, a final model is proposed which uses random forest regression model for the purpose of prediction of total score on the basis of factors like cumulative runs, cumulative strike rate, cumulative overs and wickets left. It was observed that the error in the random forest regression model is about one-fourth of the error in current run rate model. In future, the focus will be to improve the accuracy for the model or even use another methodology to get better results.

REFERENCES

.

[1] NarasimhaMurty, M.; Susheela Devi, V. (2011). Pattern Recognition: An Algorithmic Approach.

[2] Seamus Hogan (2012) Cricket and the Wasp: Shameless self promotion (Wonkish). K. Raj and P. Padma. Application of association rule mining: A case study on team India. In International Conference on Computer Communication and Informatics (ICCCI), pages 1{6,2013.

[3] T. B. Swartz, P. S. Gill, and S. Muthukumarana. Modelling and simulation for one-day cricket. Canadian Journal of Statistics, 37(2):143{160, 2009.

[4] A. Kaluarachchi and A. Varde. CricAI: A classification based tool to predict the outcome in ODI cricket.In 5th International Conference on Information and Automation for Sustainability, pages 250{255, 2010.

[5] Bhandari, E. Colet, and J. Parker. Advanced Scout:Data mining and knowledge discovery in NBA data.Data Mining and Knowledge Discovery, 1(1):121{125,1997.

[6] D. Lutz. A cluster analysis of NBA players. In MITSloan Sports Analytics Conference, 2012.

[7] S. Luckner, J. Schroder, and C. Slamka. On the forecast accuracy of sports prediction markets. In Negotiation, Auctions, and Market Engineering, International Seminar, Dagstuhl Castle, volume 2, pages 227{234,2008.

[8] G. Gartheeban and J. Guttag. A data-driven method for in-game decision making in mlb: when to pull a starting pitcher. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '13, pages 973{979, New York, NY, USA, 2013. ACM.

[9] Achen, C. H. (1982). Interpreting and using regression. Newbury Park, CA: Sage Publications.

[10] Afifi, A. A., Kotlerman, J. B., Ettner, S. L., & Cowan, M. (2007). Methods for improving regression analysis for skewed continuous or counted responses. Annual Review of Public Health, 28, 95-111.

[11] Leo Breiman. Random Forests ,1-4

[12] Yunus Koloğlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, Burhan Özyılmaz . A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position pages {1,2,8}

[13] International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 1, pages 96-100