PANCHAKOTesSAYS
ISSN : 0976-4968
Vol-14, No.-2, November 2023

Indian Premier League Cricket Score Prediction: A Machine Learning Technique

Pradip Ghanty[1*], Ishita Kar[2], Mamta Singh[2], Pushpita Chakraborty[2], Riya Baranwal[2]

[1*] Assistant Professor, Dept. of Computer Science, Asansol Girls College, Asansol, West Bengal, India.
E-mail: pradipg@agc.ac.in

[2] B.Sc(H) Pass-out students, Dept. of Computer Science, Asansol Girls College, Asansol, India.

Abstract

Cricket is the most popular game. The Indian Premier League (IPL) is one of numerous series that are contested in the nation. It is now run by ten teams. The Indian Premier League (IPL) is a twenty-20 cricket game league played to inspire young and talented players in India. The league was conducted annually in March, April or May and has a huge fan base among India. The match depends on the luck for the team, player's performance and lot more parameters that will be taken into consideration. Cricket is a game played between two teams comprising of 11 players in each team with the result being either a win, lose or a tie. Due to bad weather conditions, the game is also washed out as it cannot be played in rain. Despite this, there is a huge interest among the spectators to do some prediction either at the start of the game or during the game. Many spectators also play betting games to win money. In this paper, a model with a mechanism for predicting score is provided. In this paper, two popular machine learning techniques linear regression and random forest regression are used to score prediction. Random Forest Classifier is utilised for high accuracy and stability, ensuring that the intended anticipated output is accurate.

KEYWORDS: Machine Learning, Linear Regression, Random Forest Classifier, IPL Score Prediction

INTRODUCTION

Cricket is the second-most popular sport (behind soccer/football with 3.5 billion followers). But in India cricket is the most popular game. The Indian Premier League (IPL) is one of numerous series that are contested in the nation. Different approaches have been taken by researchers to predict the score or wining team using machine learning techniques [1-4]. In this article, a machine learning model is presented that can accurately predict the next over score of a team in an IPL cricket match based on various features such as venue, batting team, bowling team, striker, bowler, etc. By solving this problem, the IPL score predictor aims to enhance the understanding of cricket matches and provide valuable predictions that can be used for various purposes [2].

The IPL score predictor model can be utilized in various ways:

1. Match Analysis: The predictor can be used to analyse the performance of teams and players during IPL matches. It can provide insights into the factors that contribute to high or low scores, identify key players, and highlight important match conditions.
2. Strategy Planning: Team management and coaches can utilize the predictions to formulate effective strategies for batting, bowling, and fielding. By understanding the expected score, teams can plan their approach accordingly and make informed decisions about team composition, batting order, bowling strategies, and field placements.
3. Live Match Predictions: During live matches, the score predictor can provide real-time predictions of the final score based on the current match conditions. This can engage fans, commentators, and analysts in discussions and predictions about the match outcome.
4. Fantasy Cricket: Fantasy cricket platforms can incorporate the score predictor to enhance the user experience. It can help participants in making informed decisions while selecting players for their fantasy teams and predicting the scores they are likely to achieve.
5. Player Performance Evaluation: The predictor can assist in evaluating the performance of individual players by predicting their potential scores based on historical data. This can be useful for team selection, player auctions, and performance analysis.
6. Broadcast Enhancements: Television broadcasters and online streaming platforms can leverage the score predictor to enhance their coverage of IPL matches. It can be used to display real-time score predictions, statistics, and analysis to engage viewers and provide a more immersive viewing experience.

METHODS

Different machine learning algorithms have been used for live score and wining prediction [5]. Previous studies [2-6] of IPL score prediction suggest us to use Linear Regression and Random Forest Regression as the machine learning techniques for IPL score prediction.

Linear Regression 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,

$$Y = M0 + \Sigma Mi * Xi + e \quad (1)$$

In the equation (1), M0 is the intercept and Mi is the co-efficient that is learned during the model training time. Xi is the input variable and i denote their number and e is simply an error function.

Random Forest is an ensemble technique that is used to perform regression and classification tasks with the use of multiple decision trees. Ensemble techniques combine results of various machine learning models and give the best accurate prediction of any individual model. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

EXPERIMENTS

In Figure 1, we presented the steps to build a model that can consider various parameters that contribute to the score prediction.

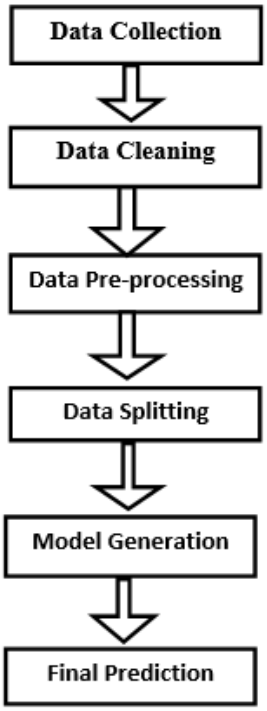


Figure 1: Block diagram of IPL Score Predictor

Data Collection: We will be taking the dataset from the datasets available on Kaggle [6]. The dataset will be taken in the CSV format. The data Collected from the website will be cleaned in the next step.

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. 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 over. 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.

Data Pre-processing: After cleaning the data, we will need our data to be pre-processed. In the data pre-processing 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.

Data Splitting: After data pre-processing, 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.

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

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 (trained model) we will be predicting a range of the score i.e., from lower bound to the upper bound.

DATASET

The dataset is a crucial component of the predictor as it provides the historical data required for training and evaluating the score prediction model. We have imported the IPL Data (2008-2019) from Kaggle [7]. The dataset contains several columns that capture various aspects of the IPL matches. In Table 1, an overview of the columns present in the dataset is shown. The samples from 2008 to 2015 are used to train the models and remaining samples (2016 to 2019) are used to evaluate the models.

The dataset contains data for multiple IPL matches, capturing various events, such as runs scored, wickets taken, and bowling and batting details. This data is utilized to train the score prediction model, allowing it to learn patterns and relationships between the input features and the target variable (score).The dataset is preprocessed and features are encoded before being used for training and testing the model. It

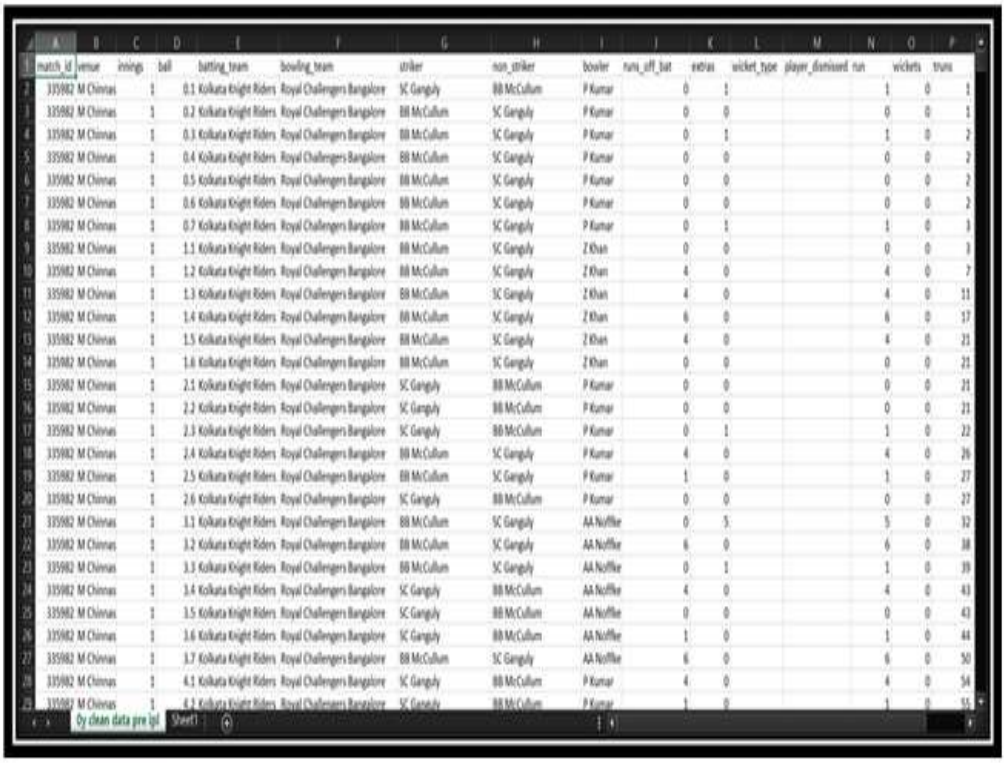
undergoes steps such as data cleaning, handling missing values, and label encoding to convert categorical features into numerical representations.

Overall, the dataset serves as a valuable resource for building an IPL score prediction model, providing the necessary information to train and evaluate the model's performance. In Figure 2, a snapshot of dataset is shown.

Table 1: List of attributes in IPL Data

Sl.	Attributes	Values
1	**venue**	Represents the venue or stadium where the match took place.
2	**innings**	Indicates the innings number (1st innings or 2nd innings)
3	**batting_team**	Refers to the team that was batting during the innings
4	**bowling_team**	Refers to the team that was bowling during the innings
5	**striker**	Represents the batsman who was facing the ball
6	**non_striker**	Represents the batsman at the non-striker's end
7	**bowler**	Refers to the bowler who was bowling the ball
8	**runs_off_bat**	Represents the number of runs scored by the batsman
9	**extra_runs**	Represents the total extra runs (wide runs, bye runs, leg byeruns, noball runs, penalty runs) conceded by the bowling team.
10	**wicket_type**	Indicates the type of dismissal if a batsman got dismissed (e.g., caught, bowled, run-out, etc.).
11	**player_dismissed**	Indicates the batsman who got dismissed, if any
12	**total_runs**	Represents the total runs scored in the particular ball

Figure 2: A snapshot of used dataset



match_id	venue	innings	ball	batting_team	bowling_team	striker	non_striker	bowler	runs_off_bat	extras	wicket_type	player_dismissed	run	wickets	truns
335982	M Chinnas	1	0.1	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	P Kumar	0	1			1	0	1
335982	M Chinnas	1	0.2	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	P Kumar	0	0			0	0	1
335982	M Chinnas	1	0.3	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	P Kumar	0	1			1	0	2
335982	M Chinnas	1	0.4	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	P Kumar	0	0			0	0	2
335982	M Chinnas	1	0.5	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	P Kumar	0	0			0	0	2
335982	M Chinnas	1	0.6	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	P Kumar	0	0			0	0	2
335982	M Chinnas	1	0.7	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	P Kumar	0	1			1	0	3
335982	M Chinnas	1	1.1	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	Z Khan	0	0			0	0	3
335982	M Chinnas	1	1.2	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	Z Khan	4	0			4	0	7
335982	M Chinnas	1	1.3	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	Z Khan	4	0			4	0	11
335982	M Chinnas	1	1.4	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	Z Khan	6	0			6	0	17
335982	M Chinnas	1	1.5	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	Z Khan	4	0			4	0	21
335982	M Chinnas	1	1.6	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	Z Khan	0	0			0	0	21
335982	M Chinnas	1	2.1	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	P Kumar	0	0			0	0	21
335982	M Chinnas	1	2.2	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	P Kumar	0	0			0	0	21
335982	M Chinnas	1	2.3	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	P Kumar	0	1			1	0	22
335982	M Chinnas	1	2.4	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	P Kumar	4	0			4	0	26
335982	M Chinnas	1	2.5	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	P Kumar	1	0			1	0	27
335982	M Chinnas	1	2.6	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	P Kumar	0	0			0	0	27
335982	M Chinnas	1	3.1	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	AA Noffke	0	5			5	0	32
335982	M Chinnas	1	3.2	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	AA Noffke	6	0			6	0	38
335982	M Chinnas	1	3.3	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	AA Noffke	0	1			1	0	39
335982	M Chinnas	1	3.4	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	AA Noffke	4	0			4	0	43
335982	M Chinnas	1	3.5	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	AA Noffke	0	0			0	0	43
335982	M Chinnas	1	3.6	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	AA Noffke	1	0			1	0	44
335982	M Chinnas	1	3.7	Kolkata Knight Riders	Royal Challengers Bangalore	BB McCullum	SC Ganguly	AA Noffke	6	0			6	0	50
335982	M Chinnas	1	4.1	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	P Kumar	4	0			4	0	54
335982	M Chinnas	1	4.2	Kolkata Knight Riders	Royal Challengers Bangalore	SC Ganguly	BB McCullum	P Kumar	1	0			1	0	55

RESULTS AND DISCUSSION

The training of linear regression model is performed with python machine learning library. Then from test data we plot the actual vs predicted values and shown in Figure 3. This figure helps visualize the relationship between the actual and predicted value.

Figure 3: Actual vs Predicted values of test data for linear regression model

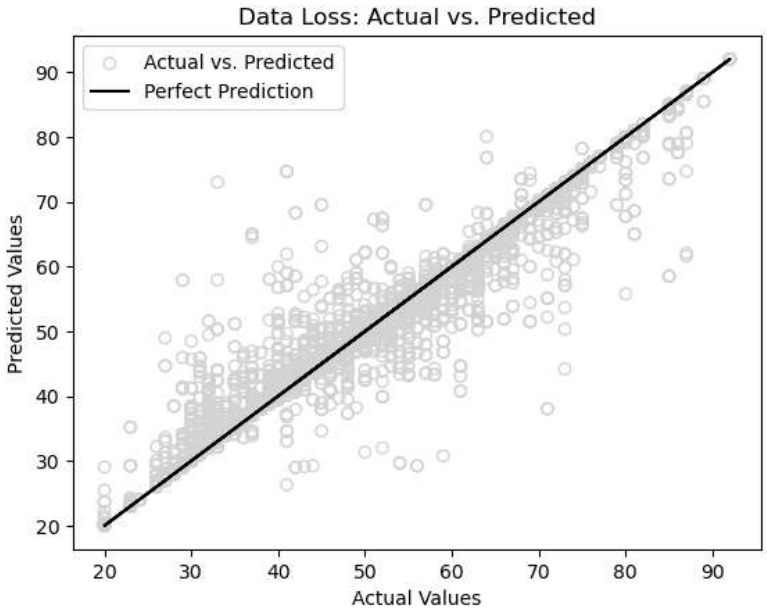
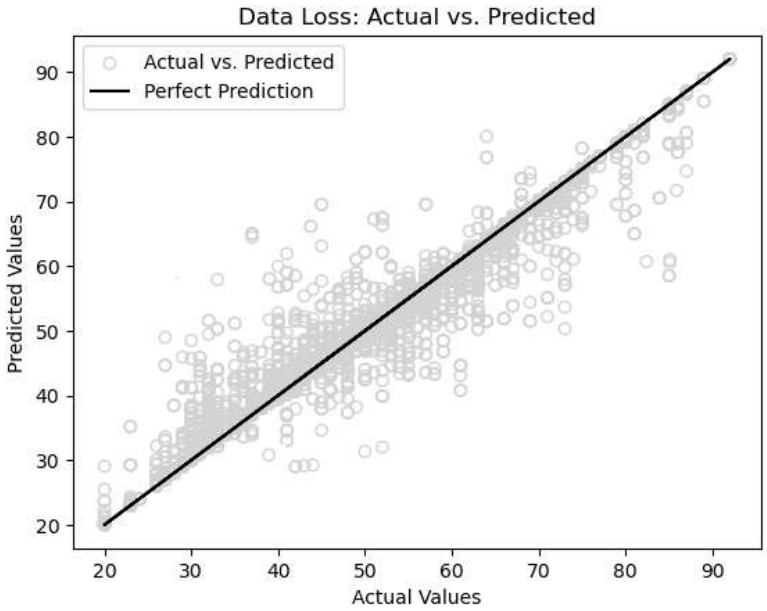


Figure 4: Actual vs Predicted values of test data for random forest regression model



From Figure 3, it is observed only few test samples (around 5-6%) produces high test errors. These samples may be outliers. To further investigate this we used random forest regression. The actual vs predicted values of test data are shown in Figure 4. Here we observed that testing error drastically reduced with random forest regression model. In Table 2, we have illustrated error obtained for few data points using linear regression and random forest regression. The overall performance of random forest regression is better than linear regression. It is also observed in the previous study [6]. Overall performance of linear regression and random forest regression is presented here in term of Mean Squared Error (MSE). The MSE is defined as Mean or Average of the square of the difference between actual and predicted values. The MSE of test data for linear regression and random forest regression are 9.57 and 3.04 respectively. Here also we observed that random forest regression is better method than linear regression for IPL score prediction.

Table 2: Errors obtained by linear and random forest regression for few IPL test data

Actual Value	Linear Regression		Random Forest Regression	
	Predicted Value	Error	Predicted Value	Error
46	41.94	4.06	43.68	2.32
50	47.02	2.98	49.07	0.93
54	58.02	-4.02	56.06	-2.06
51	53.12	-2.12	52.20	-1.20
64	67.53	-3.53	65.84	-1.84
47	45.00	2.00	46.06	0.94
48	49.56	-1.56	49.88	-1.88
34	37.46	-3.46	35.62	-1.62
48	51.00	-3.00	49.78	-1.78
50	53.00	-3.00	51.57	-1.57
48	46.56	1.44	47.32	0.68

CONCLUSIONS AND FUTURE SCOPE

In conclusion, the IPL Score Predictor study aimed to develop a machine learning model that can predict the score of an IPL cricket match based on various input features such as batting team, bowling team, venue, overs, runs, wickets, runs in the last 5 overs, and wickets in the last 5 overs. The study involved data pre-processing, feature engineering, model training and evaluation. This paper successfully implemented a Random Forest Regression model to predict the score. The model achieved a reasonable level of accuracy (in terms of MSE) in predicting the scores based on the given input features. The predictions were compared with the actual scores, and the results indicated a close correlation between the predicted and actual scores. This work focuses on exploring IPL data and presenting its insights as graphical representation and comparative analysis. By making use of this, Indian Premier League and the fan followers can take decisions on the team's performance [2].

Although the IPL Score Predictor work has achieved its primary objective of score prediction, there are several avenues for future improvements and enhancements like Incorporating real-time data, Fine-tuning the model, Feature selection and engineering, Multi model ensemble approach, Predicting other match outcomes.

Overall, the IPL Score Predictor study has demonstrated the potential of machine learning techniques in predicting cricket scores. With further enhancements and developments, it can become a valuable tool for cricket enthusiasts, analysts, and betting platforms in making informed decisions related to IPL matches.

REFERENCES:

[1] Nikhil Dhonge, Shraddha Dhole, Nikita Wavre, Mandar Pardakhe, Amit Nagarale, “IPL CRICKET SCORE AND WINNING PREDICTION USING MACHINE LEARNING TECHNIQUES”, International Research Journal of Modernization in Engineering Technology and Science, Vol. 03, Issue 05, May-2021, pp. 1723-1730.

[2] G. Sudhamathy and G. Raja Meenakshi, “PREDICTION ON IPL DATA USING MACHINE LEARNING TECHNIQUES IN R PACKAGE”, ICTACT JOURNAL ON SOFT COMPUTING, Vol. 11, Issue 01, OCTOBER 2020, pp. 2199-2204.

[3] Agrawal, S. P. Singh and J. K. Sharma, "Predicting Results of Indian Premier League T-20 Matches using Machine Learning", 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 67-71, doi: 10.1109/CSNT.2018.8820235.

[4] H. Barot, A. Kothari, P. Bide, B. Ahir and R. Kankaria, "Analysis and Prediction for the Indian Premier League," International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9153972.

[5] Rameshwari Lokhande, P. M. Chawan "Live Cricket Score and Winning Prediction", International Journal of Trend in Research and Development (IJTRD), Vol. 5, Issue1 , February 2018, pp. 30-32.

[6] K C Srikantaiah, Aryan Khetan, Baibhav Kumar, Divy Tolani, Harshal Patel, “Prediction of IPL Match Outcome Using Machine Learning Techniques”, Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021), pp. 399-406.

[7] Kaggle IPL Dataset, https://www.kaggle.com/datasets/ramjidoolla/ipl-data-set.