International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-5, June 2021

Cricket Score Forecasting using Neural Networks

Prateek Gupta, Navya Sanjna Joshi, Raghuvansh Tahlan, Darpan Gupta, Saakshi Agrawal

Abstract: Today, Sports is not what it used to be a decade ago. Technologies like Machine Learning and Artificial Intelligence have dominated it. Now there are sensors in all types of sports equipment like cricket bats, stumps, flannels, etc., which analyse the data and provide analytics, which may or may not be helpful, but we, as spectators, thoroughly enjoy the game. The terms such as Cric-Science (Cricket + Data Science) and Cricket Analytics are the fruit of ML/AI. In the last decade alone, cricket has witnessed many changes, such as the addition of a new format like T10, which is yet to be recognised by ICC, along with the introduction of many other international leagues such as IPL, BBL, PSL, CPL, apart from the widely recognised formats like Test Match, One day International and T20. With so much cricket played, the data generated is also massive. But even with these technological advancements, run rate is conventionally used to predict a team's score in the upcoming overs. So, in this research paper, we aim to predict a team's score using Neural Network by using the data from past balls.

Keywords:Cricket analytics, Cricket Score Prediction, LSTM, Neural Network, Sports Analytics, RNN, Cricket Score Forecasting

I. INTRODUCTION

An extensive online search produced very few articles related to players' performance prediction in cricket. A minimal number of researchers have studied the performance of cricket players. Muthu Swamy and Lam [1] predicted the performance of Indian bowlers against seven international teams against which the Indian cricket team plays most frequently. For this, they measured the performances of individual players of each team. Each player's performance varies with factors like the team he is playing against and the ground at which the match is being played. Therefore, it is essential to select the right players that can perform the best in each match. A player's performance depends on several factors like his current form, performance against a particular team, performance at a specific venue, etc.

Manuscript received on June 13, 2021.
Revised Manuscript received on June 20, 2021.
Manuscript published on June 30, 2021.
* Correspondence Author
Prateek Gupta*, Department of Computer Science & Engineering, Dr Akhilesh Das Gupta Institute of Technology & Management, New Delhi, India.Email: guptaprateek343@gmail.com
Navya Sanjna Joshi, Department of Computer Science & Engineering, Dr Akhilesh Das Gupta Institute of Technology & Management, New Delhi, India. Email: joshinavya3@gmail.com
Raghuvansh Tahlan, Department of Computer Science & Engineering, Dr Akhilesh Das Gupta Institute of Technology & Management, New Delhi, India. Email: raghuvanshtahlan@gmail.com
Darpan Gupta, Department of Computer Science & Engineering, Dr Akhilesh Das Gupta Institute of Technology & Management, New Delhi, India. Email: darpannguptaa@gmail.com
Ms. Saakshi Agrawal, Assistant Professor Department of Computer Science & Engineering, Dr Akhilesh Das Gupta Institute of Technology & Management, New Delhi, India. Email: saakshi.agrawal@adgitmdelhi.ac.in

They developed algorithms to model the performances of batsmen and bowlers where they determine the potential of a player by examining his career performance and then his recent performances. This section discusses the literary work of other authors in the game of cricket in data analytics.

II. BASIC TERMINOLOGIES

2.1 RNN

A Recurrent Neural Network (RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other. Still, in cases like when it is required to predict the next word of a sentence, the previous words are needed, and hence there is a need to remember the preceding words. Thus, RNN [4] came into existence, which solved this issue with the help of a Hidden Layer.

2.1.1 Limitations

Gradient vanishing and exploding problems are usually faced while working with RNNs. If the largest eigenvalue is greater than 1, then gradients will explode, and if the largest gradient value is less than 1, the gradients will vanish. Training them on a larger sequence is also a challenging task.

2.2 LSTM

Long Short-Term Memory's (LSTMs) edge over conventional feed-forward neural networks and RNN in many ways. This edge is mainly because of their property to selectively remember long patterns, which is their default behaviour, not something they struggle to learn.

2.3 Bidirectional LSTM

They are an extension of LSTM where two LSTM models are trained on all input sequence, one from Left to right and the other one from right to left.

III. PROPOSED SYSTEM

We have proposed a system which overcomes the major weakness of manual work which is time consuming and required man power to manually maintain the records and statistics of each player. For our research, we have curated our dataset sourced from ESPNCricket info [4]. Our dataset contains data from over a decade (2010 - 2021) filtered for T20 matches, amounting to over 2600 matches. Our dataset includes data from IPL (2010-2021), BBL (2011/12-2020/21), PSL (2015/16-2020/21), CPL (2013-2020), Lanka Premier League (2019/20), Mzansi Super League (2018/19-2019/20), Vitality Blast (2017-2020) and T20 International Matches (2010-2021).

We have used Commentary Data for this purpose which includes features such as Batting Team Name, Bowling Team Name, Batsman Name, Bowler Name, Runs scored [6] on the delivery, Date of the Match, Ground Match, Over Number, Ball of the Over, Run Rate, Innings Wide, Innings NoBall, Innings Leg Byes, etc.

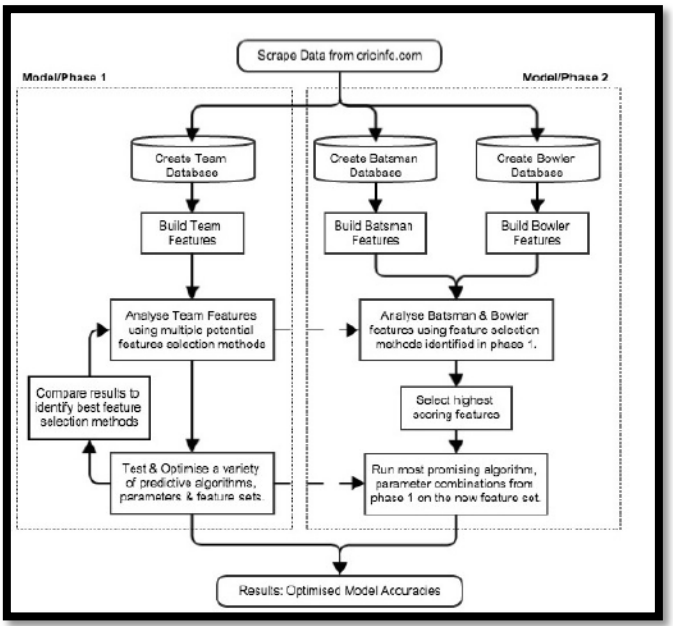


Figure 1: System Design leman

Apart from the data sourced from ESPNCricketinfo, we have also created many features such as the Number of Balls since last 4, Number of Balls since the last 6, Number of Balls since the last Wicket, Time Elapsed since Batting Team Played, Time Elapsed since Bowling Team Played, Time Elapsed since Batting Team Played on Same ground etc.

IV. FEATURE DEPLOYMENT OF MODEL

We wanted to use machine learning models like XGBoost, LightGBM, Random Forest etc., because of their convenience and readability. Still, we weren't able to use them because of the high cardinality of some of the categorical features, and they lack understanding of the context, which was important. So, we have used LSTM based Neural Network where the data of the previous 18 deliveries was used as context, and the score after 18 deliveries was our target. We created our Neural Network to illustrate the process, not to create some state-of-the-art deep learning [7] model that could be productionized. Our model uses 18 categorical and 48 numerical features. Categorical features were passed through an embedding layer and combined with numerical features using a Concatenate layer [3]. The number of embedding size used is defined by the minimum of 50 and the number of categories [1]. All the integrated features were passed through a series of LSTM and Dense layers and finally through a Dense layer containing one neuron with a suitable activation function.

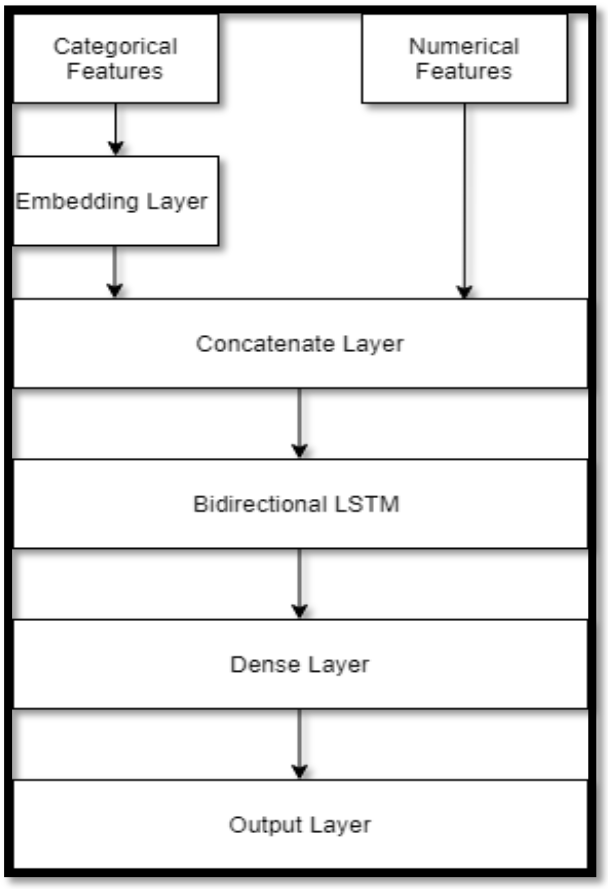


Figure 2: Neural Network Architecture

V. COMPLEXITY OF SYSTEM

As deep learning on tabular data still has its mysteries, we conducted some experiments to find optimal parameters without creating a full-blown network. Our experiments aimed to find combinations of activation function in the last layer (RELU, ELU, Linear, SoftMax, Exponential) with appropriate loss functions such as MeanSquareError, Poisson Gamma and whether scaling of independent and target variables is beneficial or not. We also conducted experiments to determine whether the learning rate from the LR range test and Cyclic Learning rate (1 Cycle Policy) would help converge the network faster or provide additional accuracy. We used the same training and testing samples in all the experiments and used the same random seed to decrease the randomness between experiments and get comparable results.

VI. MODEL BUILDING & INDUSTRIAL APPLICATIONS

However, in reality, it is probably that many extra industries will be impacted by using science as it matures. Statistical modelling has been used in sports for decades and has contributed significantly to the success on the field.

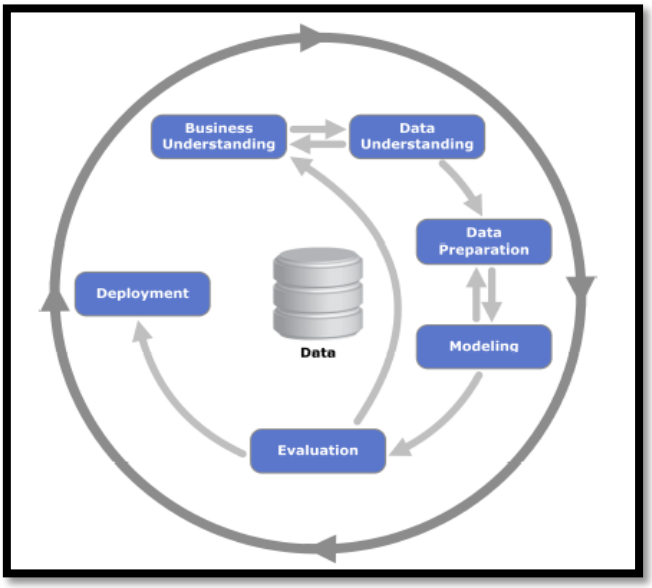


Figure 3: Model building deployment via CRISP

Various natural factors affecting the game, enormous media coverage, and a huge betting market have given strong incentives to model the game from various perspectives.

However, the complex rules governing the game, the ability of players and their performances on a given day, and various other natural parameters play an integral role in affecting the final outcome of a cricket match.

6.1 Industrial applications

Technology has transformed the whole experience of cricket for viewers [8] as well as players. Some reckon the increased use of technology as a threat as it can affect the quality of the game. However, cricket is a popular sport that cannot be kept at a bay. The sport needs to go hand-in-hand with technology.

6.1.1 SpeculurBatSense

Speculur, a Bengaluru-based sports-tech start-up collaborated with Intel to come up with a sensor named 'Specular Bat sense'. The bat sensor powered by Intel Curie technology [9] can be mounted on the top of the handle of the bat.

6.1.2 Power Bats

The idea of power bats revolves around using IoT sensors [15] to collect crucial insights on a batsman. Spektacom uses a mini sticker sensor on a cricket bat to collect data on the quality, speed, twist, and swing of the bat—to help professionals improve their game.

6.1.3 SuperStats

The website writes about SuperStats on its page as a "new language for numbers-based story-telling in cricket". The series makes use of three elements namely Smart Stats, Luck Index, and Forecaster, to predict an event in a game.

VII. FUTURE ENHANCEMENT SCOPES

Natural Language Processing has witnessed significant advancements in the last few years, such as complex Neural Network models such as BERT, OpenAI-GPT2 and GPT3, which have been nothing less than phenomenal in understanding Textual data. We wish to incorporate 'Commentary' data, which could become an essential feature for our model to predict scores.

In future, this work can be modified on the assumptions made in the starting that every player is fit and their physical training data is not available, so using the training data will increase the accuracy of the prediction. In advance of this real-time predictions on any situations can be made with the help of these calculated scores. Making strategies of the order of the batting innings or the bowling order can be sorted with these scores.

We know a team is as good as its players, but we couldn't include a feature that would indicate the Playing 11 in our dataset. So, in future, we would like to include attributes indicating the Playing 11of an innings and if they have been dismissed already or are yet to bat in the ongoing inning.

VIII. RESULTS AND OBSERVATION

Based on the mentioned and otherwise conducted experiments, we were able to make a few observations. Jason McGhee was the inspiration behind using Gamma, Poisson, Exponential and MeanSquareError as loss function [2]. Poisson distribution gave good results, while the Gamma distribution failed to perform well. We had pinned our hopes on to the LR range test and Cyclic Learning rate (1 Cycle Policy) but were left disappointed.

We conducted our experiments in three groups, one in which no transformation was used, one where we scaled only the dependent feature (Y) and one where we used scaling on both independent (X) and dependent (Y) features. The group with no transformations performed the worst, and the group with mutations only on dependent (Y) features performed the best. We also conducted experiments where we found out that Bidirectional LSTM performs better than vanilla LSTM. Two layers of Bidirectional LSTM separated by a Dense layer provided a better score than one layer.

Out of all the experiments conducted, the least MeanSqaureError was achieved when the model used the 'ELU' activation function on the output layer and MeanSquareError as a loss function. We have trained this model using Adam optimiser, Early StoppingCall-back with patience 20 and batch size of 3000, i.e., 1% of the total training samples % of the total training samples.

TRANSFORMATION	ACTIVATION FUN (OUTPUT LAYER)	LOSS FUNCTION	VALIDATION ERROR (MSE)
NO TRANSFORMATION	LINEAR	MSE	139.8
	RELU	MSE	122.8
	RELU	POISSON	117
Y - TRANSFORMED (MIN-MAX)	RELU	POISSON	90
	EXPONENTIAL	POISSON	88
	ELU	MSE	79
BOTH X AND Y TRANSFORMED (MIN-MAX)	RELU	POISSON	88.5
	EXPONENTIAL	POISSON	89.5
	ELU	MSE	99

Table 1: Outcomes of functions, MSE in hidden NN layers

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

IX. CONCLUSION

In our project, we discovered the use of LSTM based Neural Network to forecast match scores by using the information from the previous deliveries. Our research provides an alternative to Run rate-based forecasting, which is widely used but ineffective. Agreed that our process is not as easy as the former one but preliminary results sure show potential. Our most accurate model uses the 'ELU' activation function on the output layer and MeanSquareError as the loss function and outperforms the score forecasted by run rate. Our model produces RootMeanSquareError of 7.0 as compared to 18.11 generated by the run rate-based prediction. While both the models have many differences, they follow a similar error pattern achieving a minimum at around Over number 6 and elevated error past 14 Overs. While cricket has advanced by leaps and bounds, we are yet to find an accurate and efficient method for forecasting scores. Our approach shows promise, and with further research, it could unseat the conventionally followed method.

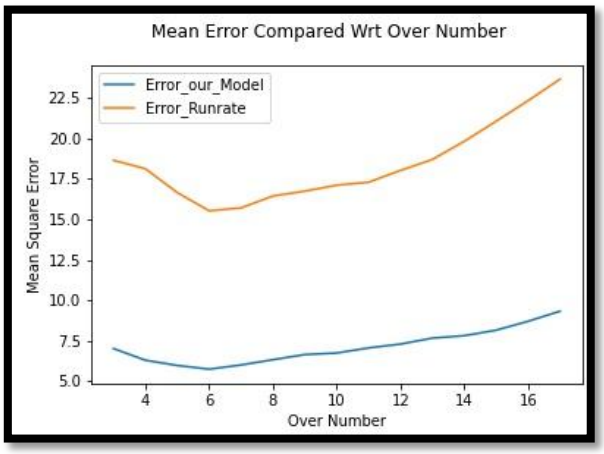


Figure 4: Comparison of MSE wrt over numbers

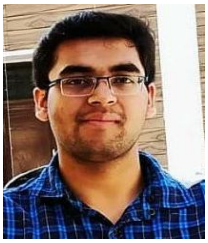
ACKNOWLEDGEMENT

This paper and the research associated would have not been possible without the magnificent support of **Ms Saakshi Agrawal (Mentor)** of Dr.Akhilesh Das Gupta Institute of Technology and Management, New Delhi. Jeremy Howard's videos were a great help in providing guidance and drawing inspiration for ideas such as Entity Embedding for the Categorical data, LR range test, Cyclic Learning Rate (1 Cycle Policy) and time series-based feature engineering. We thank our all-team members who furnished insight and understanding that extensively assisted the research. We especially thank **Dr. Saurabh Gupta** for his feedback that substantially accelerated the manuscript. We are also grateful to **Jason McGhee** for his video, wherein he shares his experience of working on tabular data, which inspired us to find the correct loss function for our Neural Network model.

REFERENCES

1. https://www.youtube.com/watch?v=WPQOkoXhdBQ&ab_channel=DataRobot
2. https://www.fast.ai/2018/04/29/categorical-embeddings/
3. https://www.youtube.com/watch?v=1-NYPQw5THU&ab_channel=JeremyHoward
4. https://www.espncricinfo.com/
5. https://www.kaggle.com/raghuvansht/cricket-scorecard-and-commentary-dataset
6. Javed, A.; Bajwa, K.B.; Malik, H.; Irtaza, A.; Mahmood, M.T. A hybrid approach for summarization of cricket videos. In Proceedings of the IEEE
7. Ananda Bandulasiri, "Predicting the Winner in One Day International Cricket" Journal of Mathematical Sciences & MathematicsEducation.
8. http://www.cricbuzz.com/
9. http://en.wikipedia.org/wiki/Cricket

AUTHORS PROFILE



Prateek Gupta, Prateek Gupta is a final year student in the Computer Science and Engineering Department in Dr. Akhilesh Das Institute of Technology & Management affiliated with GGSIPU, New Delhi. He is interested in Machine learning, Data Science along with research work applications on them & solving puzzles. He is regularly used to read new researches, upcoming technology in the field of Artificial intelligence.

Navya Sanjna Joshi, NavyaSanjna Joshi is a final year student in the Computer Science and Engineering Department in Dr Akhilesh Das Institute of Technology & Management affiliated to GGSIPU, New Delhi. She is interested and enthusiastic about researching design methodologies and formal approaches in fields of computer science like social computing, data analysis and more.

Raghuvansh Tahlan, Raghuvansh Tahlan is a final year student in the Computer Science and Engineering Department in Dr Akhilesh Das Institute of Technology & Management, affiliated to GGSIPU, New Delhi. He is interested and enthusiastic about researching the application of Data Science in Sports Analytics and Algorithmic Trading.

Darpan Gupta, Darpan Gupta is a final year student in the Computer Science and Engineering Department in Dr Akhilesh Das Institute of Technology & Management, affiliated to GGSIPU, New Delhi. He is interested and enthusiastic about solving problems related to Data Structures and Algorithms and Web Development

Ms. Saakshi Agrawal, Ms Saakshi Agrawal is working as Assistant Professor in Computer Science Engineering. Her research interest includes machine learning based real-life applications. Previous research work is based of facial expressions recognition using machine learning techniques.