Pressure Index for Cricket

Quantifying pressure to predict runs scored per ball in T20 Cricket

Rithic Kumar Nandakumar,
Student ID 20998160
rknandak@uwaterloo.ca

Prepared for CS 680 Introduction to Machine Learning Course Project
University of Waterloo, Fall 2023

Contents

1 Background

Cricket is a sport where the approach to each game depends on the context of the game. A batsman's innings with a strike rate (runs scored per 100 balls faced) of 135 can be above par in one match but might be below par in another situation where other batters have scored at a rate greater than 150. Similarly, bowlers who operate at the end of the innings (called death overs) are expected to concede more runs than those who operate in the middle phases. Hence, evaluating players using traditional statistics like strike rate, economy, runs scored or wickets taken is unfair.

This imbalance has given rise to the application of machine learning techniques to build an expected runs model that predicts run scored per ball which can be used as a baseline figure to evaluate the performance of the players. A batter who scores more runs than the expected runs predicted is said to have created a positive impact whereas the converse is true for a bowler. However, a significant challenge in the development of such metrics is the difficulty of incorporating the pressure factor into the model. Pressure on either team depends on runs scored, and wickets lost in previous balls. The expected runs for the ball will depend on the pressure accumulated as well as the current state (runs required, wickets left, balls left). Hence it is important to quantify the pressure factor to build an accurate expected runs model.

2 Related Work

Lemmer (2015) [1] devised a method to measure "choking" - a term used to denote a situation where a team in a favourable position ends up losing the match. The author makes use the product of required run rate and resources utilised (RU) based on Duckworth/Lewis table [2] to measure pressure. However, this method gives rise to unbounded pressure values and does not take the performance across previous balls into account.

Shah and Shah (2014) [3] developed a 'pressure index' based on the progress of the batting team and the runs left for the team to score. However, the work lacks clarity as many terms are left unexplained and hence cannot be reproduced.

Bhattacharjee and Lemmer (2016) [4] defined pressure index (PI_3) based on the progress of the batting team as well as the resources spent by the team.

$$PI_3 = (\frac{CRRR}{IRRR}) * 1/2[exp(RU/100) + exp(\sum w_i/11)] \quad (1)$$

where

CRRR = Current Required Run Rate

IRRR = Initial Required Run Rate

RU = Resources Utilised

$\sum w_i$ = Sum of weights of the wickets lost

where wicket weights are assigned based on work done by Lemmer (2005) for ODI and Test Matches [5]

Saikia, Bhattacharjee and Mukherjee (2019) [6] update equation 1 for T20 matches and also introduce an additional pressure index that restricts the pressure values from increasing abruptly. However, this method relies on knowing the entire values of the CRRR across the match to stabilise the values and hence cannot be utilised for prediction.

The work presented here will make use of equation 1 for T20 matches as the baseline for evaluating the proposed model.

3 Proposed Work

This project aims to develop a methodology for quantifying pressure to aid in predicting the "Expected Runs" ($xRuns$), borrowing from soccer's Expected Goals (xG) [7]. As there is no ground truth for pressure formulations, we will rely on the usefulness of the pressure formulation by using it as a feature to predict the expected runs for a ball. The formulation that achieves the least error in predicting the runs is interpreted as the most useful formulation of pressure

The proposed model has two components - a Recurrent Neural Network that quantifies pressure, and the current state variables (namely the runs left to be scored, balls consumed, wickets left and the venue of the match). The output from the Recurrent Neural Network is called the Pressure Index and is combined with the current state variable to predict the runs scored for that ball

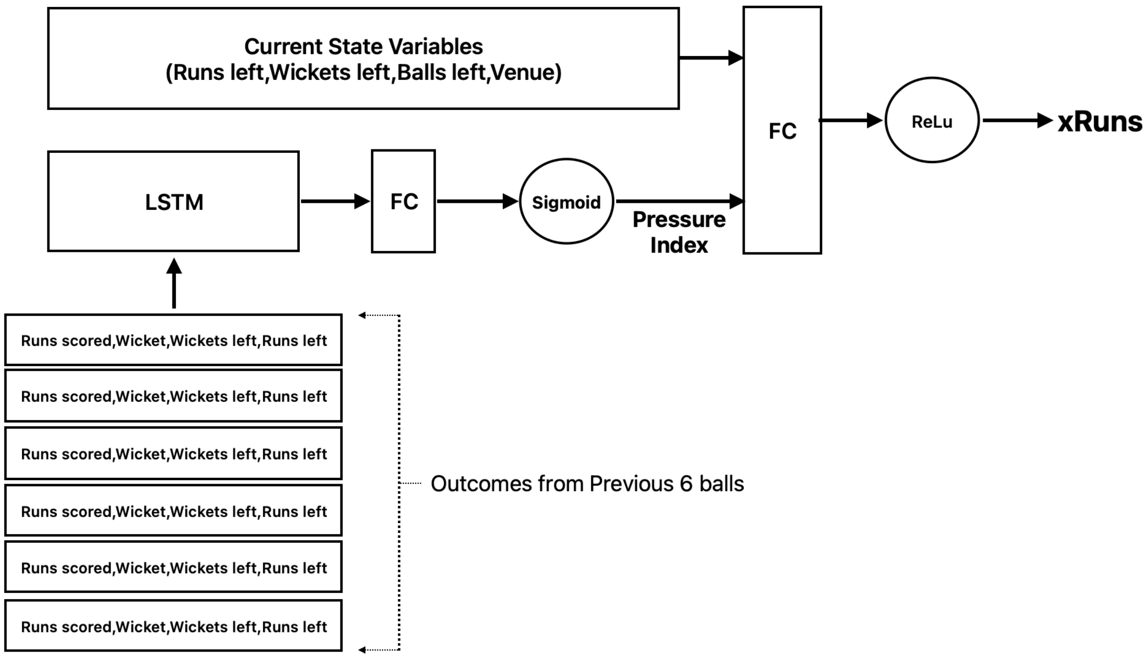

Figure 1: Model Flowchart

$$RNN(Runsscored, Wicket, Wickets\ left, Runs\ left) for\ past\ 6\ balls = Pressure\ Index \tag{2}$$

where

Runs scored = Runs scored off the bat in that ball

Wicket = binary variable denoting if a wicket fell in that ball

Wickets left = Wickets left in hand before that ball is bowled

Runs left = Runs left to be scored before that ball is bowled

$$Model(Wickets\ left, balls\ left, runs\ left, Venue, PI) = xRuns\ scored\ in\ that\ ball \tag{3}$$

Since the dynamics of each format of the game varies significantly, this project aims to develop an Expected Runs model for T20 games alone and specifically for the Indian Premier League, where the best talents across the globe compete.

4 Work Done

The CSV files for each match are obtained from cricsheet.org [8]. The training data contains matches from IPL seasons 2008 to 2021. The last two seasons (namely 2022 and 2023) are chosen as test data.

Long Short-Term Memory model is used to model the Pressure Index. We use hidden size 1 and sequence length of 6 (at each time step we feed the runs scored, wicket, wickets left and runs left for the previous 6 balls). The output of the LSTM layer is fed to a fully connected layer followed by sigmoid activation to obtain the Pressure Index. The sigmoid layer squashes the output between 0 and 1 to prevent the pressure values from rising abruptly.

The Pressure Index is combined with the current state variables via a fully connected layer to predict the runs scored. Venue variable is encoded using OneHotEncoder. ReLU activation is applied to the final outputs to keep the outputs in the positive range.

The model is trained for 500 epochs with Adam optimizer and Mean Squared Error as the loss function.

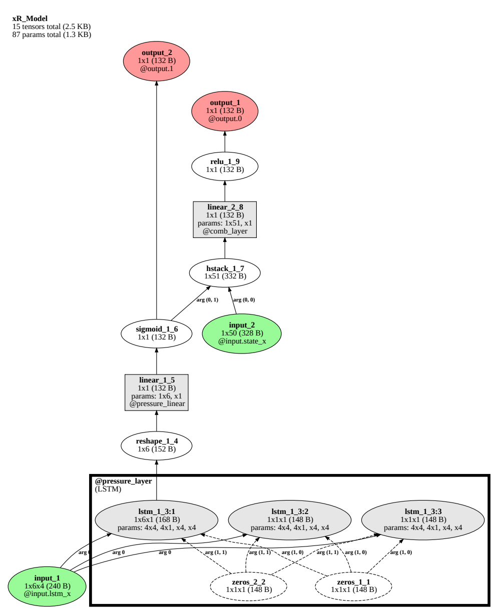

Figure 2: Model Architecture

We choose Equation (1) as the baseline Pressure Index (PI_3) and replace the Pressure

Index in the fully connected layer above. We also use a modified version of PI_3 with a sigmoid activation layer to squash the index between 0 and 1.

To test the null hypothesis that Pressure Index does not impact the quality of the expected runs (xRuns) model, Linear Regression and eXtreme Gradient Boosting (xGBoost) Regression are used to predict the xRuns with just the current state variable, ignoring the previous states.

$$Model(Wickets\ left, balls\ left, runs\ left, Venue) = xRuns\ scored\ in\ that\ ball \tag{4}$$

A naive model that outputs the average runs across each over in that venue from previous data is assumed as the worst-case prediction.

$$Average\ runs[Venue] = [a_1, a_2, ..., a_{20}] \tag{5}$$

where $a_i = Average\ runs\ per\ ball\ in\ over\ i\ in\ the\ venue$

$$Naive\ model(x_i, Venue) = a_k, \tag{6}$$

where k denotes the over ball x_i belongs to

5 Results

Mean Squared Error (MSE) is used to calculate the error between predicted values and actual runs scored by the batsman. We use seasons 2022 and 2023 as test data to evaluate the models.

$$MSE = \frac{\sum_i^n (y_i - Model(x_i))^2}{n} \tag{7}$$

where x and y denote the input features and actual runs respectively and n is the number of balls observed

We observe that our proposed model achieves the least MSE. This demonstrates the usefulness of our proposed Pressure Index compared to previous formulations.

	Model	**MSE**
Models without Pressure formulation	Naive Model	2.97
	Linear Regression	1.68e+17
	xGBoost	2.93
Models with Pressure formulation	PI_3	4.23
	PI_3 *with Sigmoid*	3.02
	Proposed Model	**2.88**

Table 1: Mean Squared Error

6 Conclusion

From the results as mentioned, it is seen that the proposed model quantifies pressure better than the existing formulations, and through the Pressure Index, the error arising while predicting runs is the least. However, there lies a significant gap in the current formulation as the model cannot be put to use on first innings data. The work presented here can be analysed and extended further to accommodate the mentioned exception and formulate Pressure Index for first innings as well.

References

[1] Hermanus H. Lemmer. "A Method to Measure Choking in Cricket". In: *International Journal of Sports Science & Coaching* 10.2-3 (2015), pp. 461–477. DOI: 10.1260/1747-9541.10.2-3.461. eprint: https://doi.org/10.1260/1747-9541.10.2-3.461. URL: https://doi.org/10.1260/1747-9541.10.2-3.461.

[2] Harsha Perera and Tim B. Swartz. "Resource Estimation in T 20 Cricket". In: 2011. URL: https://api.semanticscholar.org/CorpusID:15492943.

[3] Parag Shah and Mitesh Shah. "Pressure Index in Cricket". In: *IOSR Journal of Sports and Physical Education* 1 (2014), pp. 09–11. URL: https://api.semanticscholar.org/CorpusID:108584096.

[4] Dibyojyoti Bhattacharjee and Hermanus H Lemmer. "Quantifying the pressure on the teams batting or bowling in the second innings of limited overs cricket matches". In: *International Journal of Sports Science & Coaching* 11.5 (2016), pp. 683–692. DOI:

10.1177/1747954116667106. eprint: https://doi.org/10.1177/1747954116667106. URL: https://doi.org/10.1177/1747954116667106.

[5] Hoffie Lemmer. "A method for the comparison of the bowling performances of bowlers in a match or a series of matches". In: *South African Journal for Research in Sport, Physical Education and Recreation* 27 (Jan. 2006). DOI: 10.4314/sajrs.v27i1.25909.

[6] Hemanta Saikia, Dibyojyoti Bhattacharjee, and Diganta Mukherjee. "Performing Under Difficulty: The Magical Pressure Index". In: *Cricket Performance Management: Mathematical Formulation and Analytics*. Singapore: Springer Singapore, 2019, pp. 157–200. ISBN: 978-981-15-1354-1. DOI: 10.1007/978-981-15-1354-1_7. URL: https://doi.org/10.1007/978-981-15-1354-1_7.

[7] Hph Harm Eggels. "Expected goals in soccer:explaining match results using predictive analytics". In: 2016. URL: https://api.semanticscholar.org/CorpusID:216078599.

[8] *Cricsheet*. URL: https://cricsheet.org/.