

**ANALYSIS ON SCORE PREDICTING
IN LIMITED OVERS CRICKET MATCHES**

Uhanovitage Shakya Maduranga Senarathna

168267R

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2020

**ANALYSIS ON SCORE PREDICTING
IN LIMITED OVERS CRICKET MATCHES**

Uhanovitage Shakya Maduranga Senarathna

168267R

Dissertation submitted in partial fulfillment of the requirements for the
degree Master of Science in Computer Science and Engineering

Department of Computer Science and Engineering

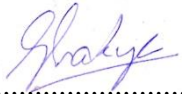
University of Moratuwa
Sri Lanka

May 2020

DECLARATION OF THE CANDIDATE & SUPERVISOR

I declare that this is my own work and this MSc dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).


.....

31-05-2020
.....

Uhanovitage Shakya Maduranga Senarathna

Date

The above candidate has carried out research for the Masters Dissertation under my supervision.

.....

.....

Dr. Amal Shehan Perera

Date

Abstract

The purpose of this research is to analyze the existing methods to predict score in one day international cricket matches and to suggest and implement a machine learning and big data based process to predict scores.

Score predicting in cricket matches is a moderately researched and published area. But target score calculation in interrupted cricket matches is heavily researched and practically in use. Since both systems use similar models, the literature review includes target score calculation models as well as score predicting models. Some researchers have tried score predicting using statistical approaches; tools like “winning and scoring predictor” (WASP) are examples for that. But the work related to these tools are not published due to the commercial value of the researches. The literature review sections contain previous work on target score calculation techniques, score predicting models and a section on application of machine learning to similar problems from other domains.

The process of preparing a dataset to build a machine learning model is discussed in detail. Match data are scraped from the web and preprocessed to build a master set of features. Then automatic feature selection algorithms are applied on the master dataset to identify the best set of features. Several representations of the same dataset with different feature set combinations are tried on a variety of machine learning algorithms. After going through several iterations, best feature set and the best machine learning model is identified.

The scope of this research is limited to score predicting in completed first innings with all 50 overs bowled. As future enhancements, the model can be extended to support all first innings as well as win percentage predictions in the second innings. A fully completed predictive model can be used as a predicting engine in news web sites. Since the research and implementation closely followed target score calculation techniques, the model can also be suggested as an alternative for current target score calculation techniques such as Duckworth Lewis Stern Method.

List of keywords - Cricket score predicting

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my supervisor Dr. Amal Shehan Perera for the guidance given to make this research and the dissertation a success. His expertise and useful suggestions helped me a lot during the research period. Further, I would like to extend my gratitude to all the academic staff for proving valuable support and material to complete this research report.

My sincere appreciation also goes to my parents, my wife and specially my brother for the support and motivation to make this research a success. Finally, I express my appreciation to all my MSc batch mates and my colleagues at Codegen International, for the support given to me to manage my MSc workload.

TABLE OF CONTENTS

Abstract	ii
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	ix
LIST OF TABLES	x
LIST OF ABBREVIATIONS	xi
1. Introduction	1
1.1 Purpose of the research	2
1.2 Research area	3
1.2.1 Why this needs a big data solution	3
1.3 Objectives	5
1.4 Scope	6
1.5 Structure of the report	6
2. literature review	8
2.1 Target score calculation	8
2.1.1 Early approaches	9
2.1.2 Duckworth Lewis Method	12
2.1.3 Jayadevan's Method	14
2.1.4 Evaluating Methods	16
2.2 Winning and Score Predictor (WASP)	16
2.2.1 Factors considered in WASP	17
2.2.2 Algorithm for WASP	17

2.3	Modelling and simulation for one-day cricket	19
2.3.1	Simulation	19
2.3.2	Modelling the distribution	21
2.4	Data mining based approaches	25
2.4.1	Problem Formulation	25
2.4.2	Features	25
2.4.3	Models	27
2.4.4	Performance	27
2.5	Result predictions in other sports	29
2.6	Evaluation techniques	29
2.7	Handling categorical data when data preparation	30
2.7.1	One Hot encoding	30
2.7.2	Ordinal encoding	30
2.7.3	Target encoding	30
2.7.4	Backward Difference encoding	31
2.7.5	Binary Coding	31
3.	Methodology	32
3.1	Introduction	32
3.2	Collecting Data	32
3.2.1	Data sources for feature extraction	33

3.3	Data Collection Process	35
3.3.1	Introduction	35
3.3.2	Tools used to web scraping	36
3.3.3	Match data extraction process	37
3.3.4	Player data extraction process	39
3.3.5	Storing the information	39
3.4	Pre-processing	39
3.4.1	Challenges faced when preparing dataset	40
3.4.2	Handling missing data	42
3.4.3	Handling categorical data	42
3.4.4	Calculating fields for target encoding	43
3.5	Predicted attribute – Score	47
3.6	Features	48
3.6.1	Match state features	49
3.6.2	Historical features	55
3.6.3	Time related features	59
3.7	Tools used	60
3.7.1	Web scraping tools	60
3.7.2	Visualizing tools	60
3.7.3	Preprocessing tools	61

3.7.4	Model building tools	61
3.7.5	Tensorflow	61
3.8	Building the model	62
3.8.1	Algorithms used	62
3.8.2	Training and test data	64
4.	Results and Evaluation	65
4.1	Evaluation matrices	65
4.1.1	percentage Error	65
4.1.2	Other standard measures	65
4.2	Identifying most productive features	66
4.2.1	By weights of a linear regression model	66
4.2.2	Feature selection algorithms	68
4.3	Modeling Algorithms used	69
4.4	Most productive feature identification process	70
4.4.1	Programmatically finding the most productive combination	70
4.4.2	Reasons for using a global search to select the best combination of features and modeling algorithm	72
4.4.3	Distribution of percentage error	74
4.5	Comparison with the DL method	74
4.6	Result conclusion	75
5.	Future work and challenges	76

5.1	Consider uncompleted innings	76
5.2	Predicting at multiple points	76
5.3	Second innings prediction	76
5.4	Real time match prediction	77
6.	Conclusion	78
	References	80
	Appendix	85

LIST OF FIGURES

Figure 2.1	The Clark curves [9]	12
Figure 2.2	Normal and target score curves of VJD method [11]	15
Figure 2.3	Factors considered in WASP [13]	18
Figure 2.4	Pseudo code used in by Swarts et al. [14]	21
Figure 2.5	probability distribution function for the model [14]	23
Figure 2.6	Performance of the non-home run model [17]	28
Figure 2.7	Performance of home run model [17]	28
Figure 3.1	Robots.txt file of espnricinfo.com [21]	34
Figure 3.2	Score distribution	48
Figure 3.3	Score at 30 overs vs final score	49
Figure 3.4	Correlation between final score and scores at 5, 10, 15, 20, 25, 30 overs	50
Figure 3.5	Number of wickets at 5, 10, 15, 20, 25, 30 overs	52
Figure 3.6	Scores and number of balls faced for the batsmen at crease at 30 overs	53
Figure 3.7	Score distribution for Home. Away, Neural matches	57
Figure 3.8	Overall average by year	60
Figure 4.1	Feature weights in Linear regression	66
Figure 4.2	Algorithm to search best combination	71
Figure 4.3	Cumulative Frequency graph of absolute errors	74

LIST OF TABLES

Table 2.1 Simplified Resource Table of the Duckworth Lewis Method [6]	13
Table 2.2 Identified situations of an innings [14]	24
Table 3.1 Batsmen with the highest batting averages	44
Table 3.2 Best bowling averages	46
Table 3.3 Team averages against other teams	47
Table 3.4 Ground averages for grounds with minimum 5 completed matches	58
Table 4.1 Performance of Predictive models for the entire feature set	69
Table 4.2 Best combination of features and models	71
Table 4.3 Comparison between the developed model and the DLS	74
Table Appendix.1 The full resource table of the Duckworth/Lewis method – standard edition [31]	85

LIST OF ABBREVIATIONS

Abbreviation	Description
ARR	Average run Rate
CDF	Cumulative Distribution Function
DLS	Duckworth Lewis Stern
DL	Duckworth Lewis
ODI	One Day International
PDF	Probability Density Function
RNN	Recurrent Neural Network
SVR	Support Vector Regression
VJD	V. Jayadevan's Method
WASP	Winning and Scoring Predictor