Cricket Match and Score Prediction for One day Internationals

Final Project Report

Submitted by

G.Sreenivaasan

Roll No: CED14I008

Degree:B.tech + M.Tech(Dual Degree) in Computer Engineering

Under the Guidance of

Dr.B.SivaSelvan

Indian Institute of Information Technology Design and Manufacturing

Kancheepuram, Melakottaiyur, Chennai-600127

May 2019

ACKNOWLEDGEMENTS

We would like to take this opportunity to thank my guide/mentor Dr. B.SivaSelvan .He has always guided in the right direction, encouraged and made me confident person in my academic career.

I would also like to thank IIITDM COE faculty Dr. N Sadagopan, Dr.V Masilamani,Dr. Noor Mahammad, Dr. J Umarani for supporting me through out my Academic life. Their inspiring/motivating lectures made a better person in my academic as well as personal life.

Bonafide Certificate

This is to certify that the titled "**Cricket Match and Score Prediction for One day Internationals**" submitted by Sreenivaasan G(CED14I008) to the Indian Institute of Information Technology Design and Manufacturing, Kancheepuram for the award of "**B.tech + M.Tech(Dual Degree) in Computer Engineering**", is a bonafide record of the project work done by him under my supervision. The contents of the project, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr.B.SivaSelvan

Project Guide
Assistant Professor
Computer Engineering
Indian Institute of Information Technology Design and Manufacturing
Kancheepuram

Place: Chennai
Date: May 2019

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [**?**] [**?**] [**?**]

Contents

List of Figures

Abstract

With the progression in innovation and in addition in sports, predicting the results, Scores of a match has turned out to be so basic. Cricket is the most popular game in Asian countries, a lot of series are played across the set of different countries during a year. Cricket is the second most watched sport in the world after soccer, and enjoys a multi-million dollar industry. This main focus of this project will be Score and Match Prediction for One Day Internationals using machine learning concepts. Currently, in One Day International (ODI) cricket matches first innings score is predicted on the basis of Current Run Rate which can be calculated as the amount of runs scored per the number of overs bowled. It does not include factors like number of wickets fallen, Number of balls left, How much the team had scored in last 5 overs, etc. Score prediction has implemented through Linear Regression and Random Forest. Random Forest Model was best with an accuracy of 0.79. Winning in Cricket depends on various factors like home crowd advantage, performances in the past,physical strength of the two teams, etc. The goal is to predict the outcome of a cricket match and improve upon the currently reported accuracy of the learners. Following learners:Logistic Regression, Naive Bayes, Decision Tree and KNN were used. Each of the models recorded major performance improvement, with Logistic Regression best with an accuracy of 0.73. However, the unpredictable rules governing the game, the capacity of players and different parameters play an essential part in influencing the ultimate result of a cricket match.

1 Introduction

Cricket is a bat-and-ball team sport first documented as being played in southern England in the 16th century. By the end of the 18th century, cricket had developed to the point where it had become the national sport of England. The expansion of the British Empire led to cricket being played overseas and by the mid-19th century the first international matches were being held. Today, the sport is played in more than 100 countries A cricket match is played involving two teams of eleven players each. The batsman looks for making runs by hitting the ball being bowled to him. The bowler on the other hand tries to get the batsman out. There are certain rules defined to get the batsman out by the bowlers or the fielders. Each batsman keeps on batting until he gets out. So, the innings of the batting team is over when either the 10 batsmen got out or the 50 overs have been bowled by the fielding team, in either of the situation the batting team now gets the chance of bowling and the bowling team gets the chance of batting. Unlike other sports, cricket stadium's size and shape is not fixed except the dimensions of the pitch and inner circle which are 22 yards and 30 yards respectively.

The cricket rules do not mention the size and the shape of the field of the stadium [1]. Pitch and outfield variations can have a substantiate effect on batting and bowling. The bounce, seam movement and spin of the ball depends on the nature of the pitch. The game is also affected by the atmospheric conditions such as altitude and weather. A unique set of playing conditions are created due to these physical differences at each venue. Depending on these set of variations a particular venue may be a batsman friendly or a bowler friendly

The game of cricket is played in three formats - Test Matches, ODIs and T20s. We focus our research on ODIs, the most popular format of the game. Game play is divided into two innings, with one team batting in the first innings and the other team bowling. Roles are reversed at the end of the innings. Which team bats first is decided by coin toss. A one day international is played with fifty overs bowled in each innings, and a twenty-twenty game is played with twenty overs for each innings.That apart, rest of the rules are the same. The team scoring the highest runs wins the game.

From [2] "one day match", so called because each match is scheduled for completion in a single day, is the most common form of limited overs cricket played on an international level. In practice, matches sometimes continue on a second day if they have been interrupted or postponed by bad weather. The main objective of a limited overs match is to produce a definite result and so a conventional draw is not possible, but matches can be undecided if the scores are tied or if bad weather prevents a result. Each

team plays one innings only and faces a limited number of overs, usually a maximum of 50 (300 deliveries).

Statistical modeling has been used in sports since decades and has contributed significantly to the success on field. 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. This presents significant challenges in predicting the accurate results of a game.

2 Problem Definition

Currently, in an ODI match the projected scores can be seen displayed at the score card during the first innings, which is basically the final score of the batting team at the end of that innings if it scores according to the current run rate or a particular rate. Run rate is defined as the amount of runs scored per the number of overs bowled. However, run rate is considered as the only criteria for calculating the final score. But there are other factors too which may affect the final score like number of wickets fallen, number of balls left, on how much scores are the current batsman batting, how much the team had scored in last 5 overs, how much the team had lost wickets in last 5 overs, the nature of the pitch, how strong is the batting and bowling team.
Cricket winning can be predicted like all other games. We need to find the best attributes or factors that influence the match outcome. The result of a cricket match depends on factors like teams playing, venue, team composition, the batting and bowling averages of the each player in the team. There are many unpredictable things that happen in a cricket game like matches being washed out due to rain, a key player getting injured before the game, players changing their teams, etc. Sometimes a key player also gets inured during the game and hence is not able to take further part in the game. All these factors do affect the prediction to some extent.

Major Goals of the Project:

- To predict the results of the cricket match by taking factors like teams playing, venue, team composition, the batting and bowling averages of the each player in the team.
- Prediction of Cricket Score by taking factors like number of wickets fallen, Number of balls left, on how much scores are the current batsman batting, how much the team had scored in last 5 overs, how much the team had lost wickets in last 5 overs.

3 Literature Review

According to [3] factors contributing to winning games are imperative, as the ultimate objective in a game is victory. The aim of this study was to identify the factors that characterize the game of cricket, and to investigate the factors that truly influence the result of a game using the data collected from the Champions Trophy cricket tournament. According to the results, this cricket tournament can be characterized using the factors of batting, bowling, and decision-making. It takes into account various factors affecting the game including home team advantage, day/night effect and toss, etc., and uses the Bayesian classifier to predict the outcome of the match. A software tool called CricAI was developed. This tool outputs the probability of victory in an ODI cricket match using input factors such as home game advantage available at the beginning of the match.

[4] embarks upon a very critical aspect that the team composition changes over time. It propose novel methods to model batsmen, bowlers and teams, using various career statistics and recent performances of the players. It propose a novel dynamic approach to reflect the changes in player combinations.

[5] focuses on the prediction of likelihood of India winning or losing in One Day International (ODI) cricket match against Australia by fitting the logistic regression model. According to ICC ODI championship rating, dated 7th August 2015, India holds 2nd position with 5875 points and 115 rating by playing 51 matches. Data from actual recent matches with five independent variables and one dependent binary logistic variable are used throughout to illustrate the implementation of this successful use of mathematical and statistical principles to the solution of a practical problem in one-day international cricket match.

[6] It estimates about how well the average batting team will do against the average bowling team under given conditions and the current state of the game. In the first-innings it estimates the additional runs that can be scored with the given number of balls and wickets remaining. In the second innings it estimates the winning probability with the given number of balls and wickets remaining, runs scored at the given situation and the target given. The estimates have been made from a dynamic programming.

4 Methodology

4.1 Datasets:Match Prediction

Cricinfo [7] is a very popular site for getting cricket related information. It contains statistics of all the matches year wise. To create data sets, data was crawled from cricinfo website. First crawled URLs of all the matches which were played in the time-span of 1990-2019.
Data set has the Following Columns:

- Team-1
- Team-2
- Stadium Name
- Date
- Toss-winner (1,2)
- First-Batting (1,2)
- Match-winner (1,2)
- Players from Team -1
- Players from Team -2

Here (1,2) in Toss-winner, First-Batting, Match-winner denotes '1' for Team-1 and '2' for Team-2.
Beautiful Soup [8] is a Python library for extracting data out of HTML and XML files very easily. Beautiful Soup was used to extract required data from each of ODI match's URL.

4.2 Feature Generation

This is the process of taking raw, unstructured data and defining features (i.e. variables) for potential use in my statistical analysis. By using the ranking of player will signify the player potential as batsman or bowler and also will describe his form at the time of match.

For extracting player ranking we used ICC Men's ODI Cricket Ranking [9] as source. In site, ranking of top 100 batsman or bowler were present at the date of match. If any batsman or bowler is not present in that top 100 we give him 101 as ranking. The intuition behind this rule is that if a player is not present in top 100 batting or bowling ranking table then he is not likely to affect the result of a match, also often it is the case that good batsman is not a good bowler and vice versa. If a player is good in both batting and bowling then he will be present in table top list most likely.

ICC Players Ranking:The player rankings are a weighted average of all a player's performances, with

recent matches weighted most heavily (so the overall effect of a good or bad performance decline over time). Each match performance is given a rating out of 1000, based on a set of predetermined criteria, and these figures averaged. This means that the maximum possible overall rating is 1000, and a player gaining an rating of 900 is seen as an exceptional achievement.

Features which will be Added:

- Bating Ranking: For batting, the performance rating is based on a combination of runs scored, the rating of the opposition bowlers, match result and comparison to the overall scores in the match.
- Bowling Ranking: A bowler gains points based on wickets taken, runs conceded and match result, with more points gained for dismissing highly rated batsmen.

So, in our final dataset each player name is replaced by two rankings, his current batting and bowling rank. Our final dataset contains following features:

- Team-1
- Team-2
- Stadium Name
- Date
- Toss-winner (1,2)
- First-Batting (1,2)
- Match-winner (1,2)
- Batting rank, Bowling Rank for each player in Playing 11 of team-1
- Batting rank, Bowling Rank for each player in Playing 11 of team-2

Now the final data set has 50 features which is on the larger side. So in order to reduce it, overall(kind of average) team's batting and bowling rankings can be used. To find the Overall Batting Strength, Bowling strength. Each ranking is subtracted from 101. And all the subtracted values are added up to the overall batting and bowling strength. Intuition behind subtracting each rank from 101 is to nullify the contribution of each player having rank greater than 100.Assuming that any player not present in say top 100 bowler list is actually a batsman and does not bowl and vice versa. 1111 comes from the fact that 11 players are present in a team so combined rank which could occur is 101 *11 = 1111.

4.3 Model Building for Prediction of Match

Predicting the Outcome of the Match:

Supervised learning algorithms is used because in Supervised learning is where you have input factors (x) and a output factor (Y) and you utilize an algorithm to learn the mapping function from the input to the output. The objective is to surmised the mapping capacity so well that when you have new input data (x) that you can predict the output factor(Y) for that data.And we do not use unsupervised learning algorithms in this paper because Unsupervised learning is the place you just have input information (X) and no comparing yield factor. Following Models will be used:

- Decision Tree: It is a decision help device that uses a tree-like chart or model of decision and their conceivable results, including possible outcomes, asset expenses, and utility.It works for both categorical and continuous input and output factors. In this procedure, we split the populace or test into at least two homogeneous sets based on most critical differentiated input factor.
- K Nearest Neighbours: It is a non-parametric method used for both classification and regression.Data points are plotted on a n-dimensional space(n is number of attributes). For an unclassified data point, its k nearest neighbours class labels are noted and the label which occurs maximum number of times is given to unclassified data point. At that point, we perform grouping by finding the hyper-plane that separate the two classes exceptionally well.
- Logistic Regression: Logistic regression is used when the response variable is categorical in nature to see a binary result (1/0, Yes/No,True/False) given an arrangement of autonomous factors. You can think of logistic regression as a special case of linear regression when the outcome factor is categorical.
- Bayes Classification:Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

75 Percentage of the data is used for training and the 25 Percentage of the data is used for testing. The splitting of the data is been done in order to find out the accuracy of the algorithm. The accuracy of the algorithms are also obtained in order to compare the algorithms to check which algorithm is more efficient.

4.4 Accuracy of the Models

Implementation various ML classification algorithm using Sklearn package [10]. Cross validation techniques was used to divide our data-set into 5 parts and calculated the Accuracy of the model. In

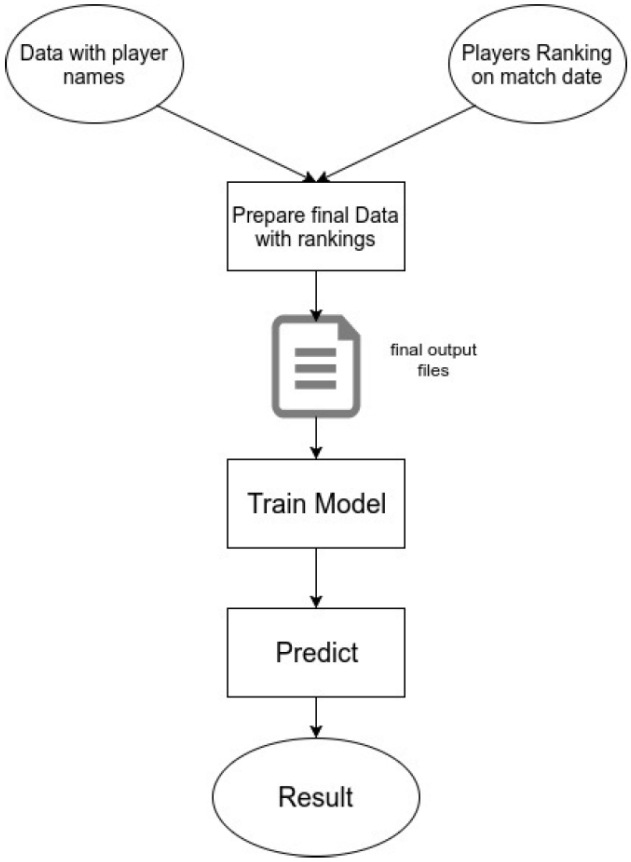

Figure 1: Workflow for Match Prediction

[11],predicting the outcome of an ODI cricket match was done using the statistics of 366 matches, they have followed a Team composition based approach and have modeled players using career statistics as well as the recent performances of a player. k-Nearest Neighbor (kNN) algorithm yields better results with an accuracy of 0.68. The drawback over here is Modeling of the bowlers since they did not have the data to model the Bowlers. Using the ICC Ranking my model had overcome this limitation and since better results. Below are list of Models and their Accuracy.

Techniques	Accuracy
Decision Tree	0.64
K-Neighbour Classifier	0.62
Logistic Regression	0.73
Bayes Classification	0.69

4.5 Datasets:Score Prediction

Cricsheet [12] was used to get ball by ball details of matches.It had the data in YAML format which was then converted to CSV format.

Data set has the Following Columns:

- Mid: Each match is given a unique number
- Date
- Stadium Name
- bat-team: Batting team name
- bowl-team: Bowling team name
- batsman: Batsman name who faced that ball
- bowler: Bowler who bowled that ball
- runs: Total runs scored by team at that instance
- wickets: Total wickets fallen at that instance
- overs: Total overs bowled at that instance
- runs-last-5: Total runs scored in last 5 overs
- wickets-last-5: Total wickets that fell in last 5 overs
- striker: max(runs scored by striker, runs scored by non-striker)
- non-striker: min(runs scored by striker, runs scored by non-striker)
- total: Total runs scored by batting team after first innings

4.6 Feature Generation

This is the process of taking raw, unstructured data and defining features (i.e. variables) for potential use in my statistical analysis.

Following data were used as Features:

- Runs: Total runs scored by team at that instance
- wickets: Total wickets fallen at that instance
- overs: Total overs bowled at that instance
- striker: max(runs scored by striker, runs scored by non-striker)
- non-striker: min(runs scored by striker, runs scored by non-striker)

4.7 Model Building for Prediction of Score

Predicting the Score:

Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.
Following are the Supervised Learning Models is used here.

- Linear Regression: Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other.
- Random Forest: It is a supervised learning algorithm. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. It can be used for both classification and regression tasks.

75 Percentage of the data is used for training and the 25 Percentage of the data is used for testing. The splitting of the data is been done in order to find out the accuracy of the algorithm. The accuracy of the algorithms are also obtained in order to compare the algorithms to check which algorithm is more efficient.

4.8 Accuracy of the Models

Implementation various ML classification algorithm using Sklearn package . R-sqaured is a statistic that will give some information about the goodness of fit of a model. In regression, the R-squared coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R-squared value of 1 indicates that the regression predictions perfectly fit the data. Here the Accuracy is based on R-squared value. Below are list of Models and their Accuracy.

Techniques	Accuracy
Linear Regression	0.65
Random Forest Regression	0.79

5 Conclusion and Future Work

As Cricket is a game which is full of uncertainties, still it depends on various factors mostly on players and some other conditions. Developing a model which would predict a match's output with accuracy around 0.90 or greater is very unrealistic. Here, I have used data of 29 years then our model predicted output with accuracy of 0.73, which is quite good given the nature of the game. I have learnt so many things in Machine Learning area in order to complete this project. I have got to know various libraries which provides implementations of various algorithms which would have been very time consuming to implement from scratch. The novelty of this approach lies in addressing the problem as a dynamic one, and using the participating players as the key feature in predicting the winner of the match. Prediction of Cricket Score at a given Instance was also successfully implemented with an accuracy of 0.79.

Future Work to be Carried out:

- To build a model which will assist the team management, Captain, Coach for the Recommendation of the team considering factors like Venue,opponent, Player Statistics, Tournament.
- To build a User Interface Tool which will Display all the results of the Prediction of any cricket ODI match given the I/p's as team, Opponent, Toss, Venue, Day/Night.
- Extend the features like Weather condition, Nature of the pitch and improve the Accuracy of the existing Models.

References

[1] "Laws of cricket. https://www.lords.org/mcc/all-laws," 2017.

[2] "Odi information webpage available at https://en.wikipedia.org/wiki/one_day_international,"

[3] A. Kaluarachchiand and S. Aparna, "Cricai: A classification based tool to predict the outcome in odi cricket," *Fifth International Conference on Information and Automation for Sustainability*, 2010.

[4] . G. Jhawar and V. Pudi, "Predicting the outcome of odi cricket matches: A team composition based approach," *ECML-PKDD*, 2016.

[5] A. B. M. Pattnaik, "Fitting of logistic regression model for prediction of likelihood of india winning or losing in cricket match," *International Conference on Management and Information Systems*, 2015.

[6] S. Brooker and S. Hogan, "Winning and score predicting (wasp)," *International Conference on Recent Innovations in Science, Engineering Technology*, 2017.

[7] "Crickinfo webpage available at http://www.http://www.espncricinfo. com,"

[8] "Beautiful soup documentation webpage available at https://www.crummy.com/software/beautifulsoup/bs4/doc/

[9] "Icc men's odi cricket webpage available at http://www.relianceiccrankings.com/mensodi.php,"

[10] S. Viswanadha and M. G. Jhawar, "Dynamic winner prediction in twenty20 cricket: Based on relative team strengths," *MLSA@PKDD/ECML*, 2017.

[11] "Wikipedia on the icc cricket world cup 2019 https://en.wikipedia.org/wiki/2019_cricket_world_cup,"

[12] "Ball by ball data available at https://cricsheet.org/downloads/,"

[13] P. Shah, "Predicting outcome of live cricket match using duckworth - lewis par score," *International Journal of Latest Technology in Engineering, Management Applied Science (IJLTEMAS)*, 2017.

[14] M. Bailey and S. Clarke, "Predicting the match outcome in one day international cricket matches, while the game is in progress," *The 8th Australasian Conference on Mathematics and Computers in Sport*, 2006.

[15] T. B. P. Swartz and D. Beaudoin, "Optimal batting orders in one-day cricket," *Computers Operations Research*, 2006.

[16] "Supervised learning webpage available at https://scikit-learn.org/stable/supervised_learning.html,"

[17] V. V. J. S. Sankaranarayanan and L. V. Lakshmanan, "Auto-play: A data mining approach to odi cricket simulation and prediction," *SIAM Conference on Data Mining*, 2014.

[18] M. Khan and R. Shah, "Role of external factors on outcome of a one day international cricket (odi) match and predictive analysis," *International Journal of Advanced Research in Computer and Communication Engineering*, 2013.

[19] S. Singh, "Measuring the performance of teams in the indian premier league," *American Journal of Operations Research*, 2011.

[20] S. Kumar and S. Roy, "Score prediction and player classification model in the game of cricket using machine learning," *INTERNATIONAL JOURNAL OF SCIENTIFIC ENGINEERING RESEARCH*, 2018.