IJARnD INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

(Volume3, Issue3)

Available online at: www.ijarnd.com

Cricket score and winning prediction using data mining

Akhil Nimmagadda[1], Nidamanuri Venkata Kalyan[2], Manigandla Venkatesh[3], Nuthi Naga Sai Teja[4], Chavali Gopi Raju[5]

[1,2,3,4]Student, Vasireddy Venkatadri Institute of Technology, Namburu, Andhra Pradesh
[5]Assistant Professor, Vasireddy Venkatadri Institute of Technology, Namburu, Andhra Pradesh

ABSTRACT

Data Mining and Machine Learning in sports analytics, is a brand-new research field in computer science with a lot of challenge. In this research the goal is to design a result prediction system for a T20 cricket match, in particular for an IPL match while the match is in progress. Different Machine Learning and statistical approach were taken to find out the best possible outcome. A very popular mathematical technique named Multiple Linear Regression is used in order to make comparison of results found. This model is very much popular in predictive modelling. Currently, in Twenty-Twenty (T20) 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, venue of the match, toss. Furthermore, in second innings there is no method to predict the outcome of the match. In this paper a model has been proposed that predicts the score in each of the innings using Multiple Variable Linear Regression along with Logistic regression and finally the winner of the match using Random Forest algorithm.

Keyword: *Data Mining, Prediction, T20, IPL, Logistic Regression, Random Forest.*

1. INTRODUCTION

Cricket is one of the most popular sports in the world, viewed by majority of world's population. It is a game played between two teams of eleven players each. With the advent of statistical modeling in sports, predicting the outcome of a game has been established as a fundamental problem. Cricket is one of the most popular team games in the world. The game of cricket is played in three formats - Test Matches, ODIs and T20s. We focus our research on T20s, the most popular format of the game. With this article, we embark on predicting the outcome of a Indian Premier League (IPL) cricket match. In an IPL season there may be a minimum of 8 to 10 teams playing and each team plays with remaining all teams for a minimum of two times. Matches are held at different venues. Initially toss plays as a crucial factor in deciding the winner of the match. Toss winning team can wish to either field or bat. The team batting first will try to pose as many runs as possible in their 20 overs in order to set a target. The team batting second need to chase the target in order to win the game with wickets in hand. For years while watching limited overs cricket, we have seen projected scores at different intervals being displayed on our television screens. Projected scores are completely based on runs scored and looking at different totals at the end of an innings, using various run rates. For example, if a team's score is 100 at the end of 10 overs. There could be four variations of projected scores:

- Current run-rate: 200
- 6 per over: 160
- 8 per over: 180
- 12 per over: 220

Considering only run rate may not yield good results since various factors might affect the score of the innings. We develop a model for T20 format games by mining existing game data which can be available from cricinfo website.

2. DESCRIPTION

Statistical modeling has been used in sports since decades and has contributed significantly to the success on the 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. To predict the outcome of ODI cricket matches, we propose an approach where we first estimate the batting and bowling potentials of the 22 players playing the match using their career statistics and active participation in recent games. We use these player potentials to render the relative dominance one team has over the other. Taking some other base features into account, namely, run rate and the venue of the match, along with the relative team strength, we adopt supervised learning algorithms to predict the winner of the match.

2.1 Algorithms

Prediction Modeling using Multiple Linear Regression: Regression is an inherently statistical technique used regularly in data mining. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, and nature of regression line is linear. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. The multiple linear regression equations is as follows:

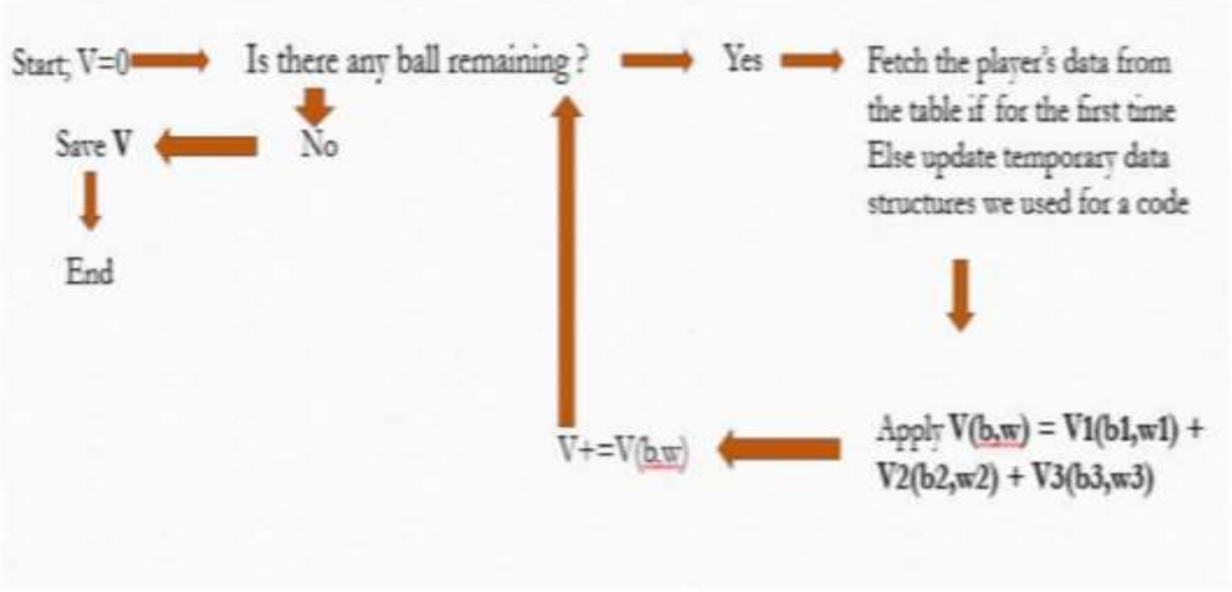
$$Y = b0 + b1X1 + b2X2 + ... + bpXp$$

where Y is the predicted or expected value of the dependent variable, X1 through Xp are p distinct independent or predictor variables, b0 is the value of Y when all of the independent variables (X1 through Xp) are equal to zero, and b1 through bp is the estimated regression coefficients. Each regression coefficient represents the change in Y relative to a one-unit change in the respective independent variable. In the multiple regression situations, b1, for example, is the change in Y relative to a one unit change in X1, holding all other independent variables constant (i.e., when the remaining independent variables are held at the same value or are fixed). Again, statistical tests can be performed to assess whether each regression coefficient is significantly different from zero.
In our model, we have taken batsmen in-crease, bowler, and wickets and run rate into consideration. The equation that we used is V(b,w)=r(b,w)+p(b,w)V(b+1,w+1)+(1-p(b,w)))V(b+1,w) Since V(b*,w)=0 where v is match variable and b would be the bowler and w is wicket. By calculating v(b,w) we would get the approximate score prediction.

Logistic Regression

It is widely used for classification problems. Logistic regression doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio. To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression. It requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square. The independent variables should not be correlated with each other i.e. no multi collinearity. However, we have the options to include interaction effects of categorical variables in the analysis and in the model. If the value of dependent variable is ordinal, then it is called as Ordinal logistic regression. If dependent variable is multi class then it is known as Multinomial Logistic regression.

2.2 Working Model


Start; V=0
Is there any ball remaining ?
Yes
Fetch the player's data from the table if for the first time Else update temporary data structures we used for a code
No
Save V
End
Apply V(b,w) = V1(b1,w1) + V2(b2,w2) + V3(b3,w3)
V+=V(b,w)

3. TOOLS AND SOFTWARE'S USED

ANACONDA

Anaconda is the installation program used by Fedora, Red Hat Enterprise Linux and some other distributions. During installation, a target computer's hardware is identified and configured and the appropriate file systems for the system's architecture are created. Finally, anaconda allows the user to install the operating system software on the target computer. Anaconda can also upgrade existing installations of earlier versions of the same distribution. After the installation is complete, you can reboot into your installed system and continue doing customization using the initial setup program. Anaconda is a fairly sophisticated installer. It supports installation from local and remote sources such as CDs and DVDs, images stored on a hard drive, NFS, HTTP, and FTP. Installation can be scripted with kickstart to provide a fully unattended installation that can be duplicated on scores of machines. It can also be run over VNC on headless machines. A variety of advanced storage devices including LVM, RAID, iSCSI, and multipath are supported from the partitioning program. Anaconda provides advanced debugging features such as remote logging, access to the python interactive debugger, and remote saving of exception dumps.

SPYDER

Spyder is the Scientific Python Development Environment:

- a powerful interactive development environment for the Python language with advanced editing, interactive testing, debugging and introspection features
- and a numerical computing environment thanks to the support of *IPython* (enhanced interactive Python interpreter) and popular Python libraries such as *NumPy* (linear algebra), *SciPy* (signal and image processing) or *matplotlib* (interactive 2D/3D plotting).

Spyder may also be used as a library providing powerful console-related widgets for your PyQt-based applications – for example, it may be used to integrate a debugging console directly in the layout of your graphical user interface.

Python Libraries

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python. Additionally, it has the broader goal of becoming **the most powerful and flexible open source data analysis / manipulation tool available in any language**.

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Scikit-learn (formerly **scikits.learn**) is a free software machine learning library for the python programming language. It features various classification, regression and clustering algorithms including support vector machines random forest, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy

4. RESULTS

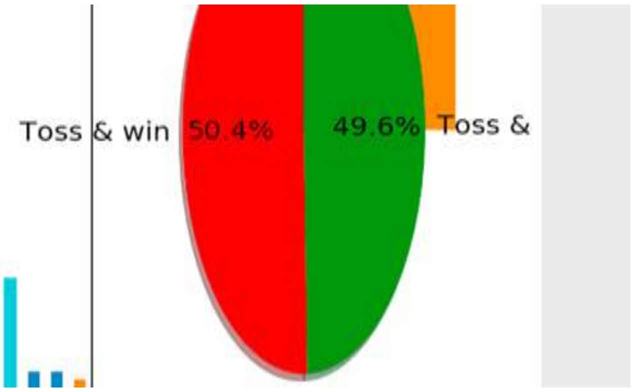


Fig -1: showing prediction based on toss

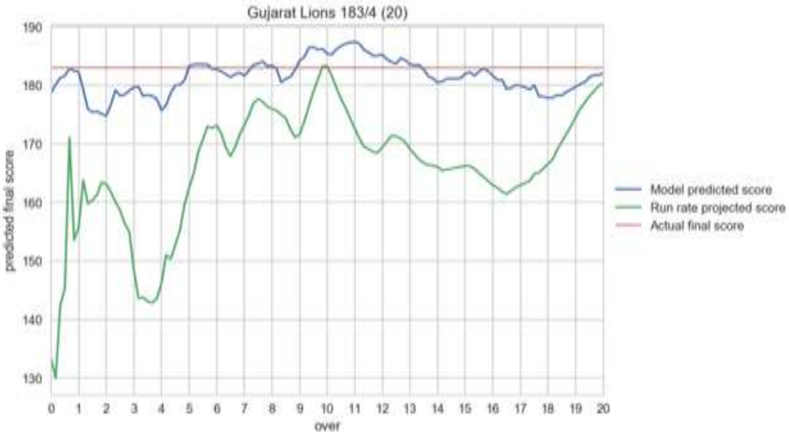


Fig -2: performance of the model

5. CONCLUSIONS

Our main goal in this paper to develop a model to predict the outcome of an ODI cricket match while the game is in progress. We used the data of previous matches played between the team in order to design our model. We have used Multiple Variable Linear Regression to design this model. Efficiency and error checking was also done in our work. Using multiple linear regression, each innings score is predicted at regular intervals and final the winner of the match. This knowledge will help us in the future to design a much more accurate prediction

6. FUTURE WORK

- As we know Machine Learning and Data Mining are developing at a rapid pace with several new techniques being developed and old techniques being modified to enhance performance, keeping this in mind our work can be expanded to incorporate new methods of classification for outcome prediction.
- More features could be added along with the ones currently considered.
- Although our study is done for ODI matches only, the however similar approach could be applied to predict outcome in other versions of Cricket matches as well.
- Classification techniques can be applied to other sports such as baseball, football as well, although
- the method of implementation might differ from one sport to another.

7. REFERENCES

[1] http://en.wikipedia.org/wiki/Cricket
[2] http://www.duckworth-lewis.com/mags/dlmethod/
[3] http://www.cricbuzz.com/
[4] Ananda Bandulasiri, “Predicting the Winner in One Day International Cricket” Journal of Mathematical Sciences & Mathematics Education.
[5] Tejinder Singh, Vishal Singla and Parteek Bhatia, “Score and Winning Prediction in Cricket through Data Mining” 8 October 2015.