ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

Role of External Factors on Outcome of a One Day International Cricket (ODI) Match and Predictive Analysis

Mehvish Khan[1], Riddhi Shah[2]

Student, Information Science and Engineering, BMS College of Engineering, Bangalore, India[1,2]

Abstract: The outcome of a One Day International (ODI) cricket match depends on various factors. This research aims to identify the factors which play a key role in predicting the outcome of an ODI cricket match and also determine the accuracy of the prediction made using the technique of data mining. In this analysis, statistical significance for various variables which could explain the outcome of an ODI cricket match are explored. Home field advantage, winning the toss, game plan (batting first or fielding first), match type (day or day & night), competing team, venue familiarity and season in which the match is played will be key features studied for the research . For purposes of model-building, three algorithms are adopted: Logistic Regression, Support Vector Machine and Naïve Bayes. Logistic regression is applied to data already obtained from previously played matches to identify which features individually or in combination with other features play a role in the prediction. SVM and Naïve Bayes Classifier are used for model training and predictive analysis. Graphical representation and confusion matrices are used to represent the various sets of models and comparative analysis is done on them. A bidding scenario is also considered to explain the decisions that can be taken after the model has been built. Effect of this decision on the cost and payoff of the model is also studied.

Keywords: Analytics, Cricket, Sports Prediction, Logistic Regression, Naïve Bayes, Support Vector Machine

I. INTRODUCTION

The outcome that takes place on a sports field are obviously heavily influenced by the ability and performance on the day of the athletes taking part; however these are not sole determinants. The influence of weather, venue conditions, and game format also plays a role in the outcome. In many sports, particularly those played outside, the ease with which player's skill and effort can translate into positive outcome can depend heavily on these conditions. One sport where this can be observed is One Day International cricket. The first official ODI match was played in 1971 between Australia and England at the Melbourne Cricket Ground.

In ODI cricket, one team bats and has a single "innings" in which it seeks to score as many runs as possible. The innings ends when the other team has bowled 300 deliveries (50 overs) to the batsman, or when 10 batsmen have been dismissed, whichever comes first. The team then change roles and the other team has an innings of 300 deliveries or 10 dismissals with which to try and achieve a higher score.

II. LITERATURE REVIEW

It is clear that the use of analytics in sports can contribute to success on the field or court, and at the ticket window thus making it highly popular. Michael Lewis' entertaining story about the use of data analysis in baseball in Moneyball: The Art of Winning an Unfair Game (Lewis 2004) is arguably the most visible account of sports analytics. Moneyball shows how any small-market Oakland A's like organizations can exploit information to gain a competitive advantage even against richer, more established organizations. Many other sports including Major League Basketball (MLB) teams have adopted the strategies used by the Oakland A's in a form suitable to them. Although Moneyball is clearly not the earliest example of applying analytics to baseball—Bill James (James 2001) [9], George Lindsey (Lindsey 1959, Lindsey 1961, Lindsey 1963), and many others preceded the work described in Moneyball-this book was the catalyst for introducing the broader sports community to the potential benefits of quantitative analysis. Moneyball became a best seller; in 2011, a movie of the same name (starring Brad Pitt) achieved considerable box office success [6,7,8].

This triggered the use of data analytics in sports and not just in the form of creating an optimal team with the least resource usage but also in the how external factors may play an impact on the field of play, predicting of match outcomes based on previously played matches, the margin by which a team might win etc. "Introduction to the Special Issue on Analytics in Sports, Part I: General Sports Applications" by Michael J. Fry, Jeffrey W. Ohlmann (2015) shows how various sports have been studied from the usage of data analytics[1]. Cricket is one such sport in which data analytics can be used in a variety of ways. In one-day international (ODI) format, for example there are an endless number of questions that can be answered with the help of data analytics. Some of the work seen were by Dyte (1998) which simulates batting outcomes between a specified test batsman and bowler using career batting and bowling averages as the key inputs without regard to the state of the match (e.g., the score, the number of wickets lost, the number of overs completed)[4]. Bailey & Clarke (2004, 2006) who worked on the how external factors play a role in determining the outcome of ODI cricket matches. Some of the more prominent factors include home ground advantage, team quality (class) and current form [2,3].

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

Other work done on predicting match outcome in ODI cricket based on external factors: Bandulasiri (2008) which uses Logistic Regression technique for exploring the statistical significance of various features and to build a model [10]. In this paper, we extend his work done by calculating match outcome probabilities using Logistic Regression and develop models for predictive analysis using Naïve Bayes and Support Vector Machine. Results are compared (graphically) to show which of the two algorithm gives a better model.

III. IMPLEMENTATION

A. *Data Description*

Data was collected manually from website, www.espncricinfo.com for all ODI matches played by Sri Lanka and India between the years 1995 to 2014. The collected data was subjected to cleaning process where some of the matches were deleted from the analysis due to certain reasons such as abundance of bad weather or when the one team was much superior to the other (ranked team playing non-ranked teams). Tied games were also deleted from the analysis. Therefore, we only study games having a clear decision. The data set was saved in comma separated format.

Data was divided into three data sets: matches played till 2007, matches played till 2014 and matches played till 2014 with additional features (venue familiarity and season) for both countries. The number of matches are given in Table I.

The datasets are named as IN2007, IN2014, IN2014+, SL2007, SL2014, SL2014+ where IN stands for India and SL stands for Sri Lanka.

TABLE I :NUMBER OF MATCHES PLAYED BY INDIA AND SRI LANKA

COUNTRY	MATCHES PLAYED TILL 2007	MATCHES PLAYED TILL 2014
INDIA	373	548
SRI LANKA	327	502

The features chosen for our analysis are Home Team advantage, Toss, Game Format (Day only or Day and night match), team which bat first and opponent team. Additional features added are: familiarity with venue and season in which match is played. We use the following formula:

Y=ax1+bx2+cx3+dx4+ex5+fx6+gx7
where Y=Winner
x1=Toss
x2=BatFirst
x3=DayNight
x4=HomeTeam
x5=Team2
x6=VenueFamiliarity and
x7=Seasons

We define,

- Winner 0: losing a match
  1: winning a match
- Toss 0: losing coin toss in match
  1: winning coin toss in match
- BatFirst 0: batting second in match
  1: batting first in match
- DayNight 0: match is day only
  1: match is day and night
- HomeTeam 0: match not played on home ground
  1: match played on home ground
- Team2 Australia, Bangladesh, England, India, NewZealand, Pakistan, Sri Lanka, South Africa, West Indies and Zimbabwe depending on Team 1 (India/Sri Lanka)
- Seasons Autumn, Monsoon, Spring, Summer, Winter depending on the country where match is held.

B. *Feature Selection*

1) Logistic Regression:

Logistic regression was used for identifying the significance of features and to determine the role played by individual as well as different combination of features. The change in the impact of these features over the years was also studied. For this part of our study the data sets IN2007, SR2007 and SR2014 were considered.

The relationship between the features Toss and Game format was studied. IN2007 data set was considered which contained a total of 373 matches (224 “Day” and 149 “Day and Night”). Table II and Table III shows the winning percentage of India for “Day” and “Day and Night” matches respectively. As seen in the tables, of the 224 “Day” only matches played by India which we considered, that nation won 58.58% the games after having lost the coin toss, however this percentage falls to 44.8% when the coin toss was won.

A logical explanation to explain this phenomenon could be that in “Day” matches, the team that wins the Toss decides their strong strategy to play with and the opponent is forced to defend against these strategies. In “Day and Night” matches, the match starts in afternoon and goes on till midnight. This gives the team winning the Toss to decide the strategy according to the changing field and weather conditions and help them win the match.

Even after winning the toss, the teams have lost a larger percentage of matches. This indicates that despite luck being on their side, they are prone to make bad decision. However, it was seen that winning the coin toss gives competitive advantage for “Day & Night” matches.

TABLE II : CLASSIFICATION OF RESULTS FOR “DAY ONLY” MATCHES PLAYED BY INDIA TILL 2007 Results

	L	W	Total
L	41(41.41%)	58(58.58%)	99
W	69(55.2%)	56(44.8%)	125
Total	110	114	224

 DOI 10.17148/IJARCCE.2015.4642

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

TABLE III : CLASSIFICATION OF RESULTS FOR "DAY AND NIGHT" MATCHES PLAYED BY INDIA TILL 2007 Results

	L	W	Total
L	40(63.49%)	23(36.5%)	63
W	40(46.51%)	46(53.48%)	86
Total	80	69	149

For the next part of our analysis the datasets SR2007 and SR2014 were considered to study the change in the effect of the feature Home Team advantage. As seen in Table IV, Home team advantage plays a major role in determining the winner of the match. This seems rational as playing at a home ground offers many benefits. Most of these are psychological in nature, such as familiarity with the playing grounds, the ability for participants to lodge in their homes rather than in a hotel, less likelihood of travel immediately prior to the game, and the support of the fans in attendance.

Table V shows the results for the ODI matches played by Sri Lanka till the year 2014. As seen, there is a drop in the odds ratio by 2, i.e., the country is only two times more likely to win a match in home ground as opposed for 4 times for the matches played till 2007. This signifies that the effect of home advantage have diminished in recent years. Possible reasons for this pattern are that there are young, enthusiastic players with international exposure, fans are willing to travel abroad to support their teams, emergence of new leagues in the recent years like Indian Premier League(IPL), Champions League T20, Big Bash League etc. which provide global exposure to all the players.

TABLE IV: LOGISTIC REGRESSION MODEL FOR SRI LANKA (2007)

	OR	p-value
Home Team	4.0285	1.792e-06***
Toss	0.8631	0.9728
DayNight	0.4944	0.02459*
BatFirst	1.4015	0.0947 .

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.0.1 ' ' 1

TABLE V: LOGISTIC REGRESSION MODEL FOR SRI LANKA (2014)

	OR	p-value
Home Team	2.4257	3.448e-05***
Toss	0.9997	0.8189
DayNight	0.5534	0.0067**
BatFirst	1.0671	0.6471

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

A. Predictive Analysis

Predictive analysis is done using three methods- hold-out method, k-fold cross validation and leave one out cross validation. All these methods used alongside algorithm help generate a confusion matrix used for predictive analysis. Among the three methods k-fold cross validation (k=10) gives the most consistent results. The confusion matrix can be compared using a certain set of parameters like sensitivity, specificity, accuracy etc. The two algorithms used for the analysis are:

1) Support Vector Machine:

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. This algorithm has been applied to the datasets SL2014+ and IN2014+ as shown below to generate confusion matrix in the form of Table VI and Table VII.

TABLE VI: CONFUSION MATRIX FOR MATCHES PLAYED BY SRI LANKA TILL 2014 FOR 10 FOLD CROSS VALIDATION (SVM)

	L	W	Total
L	170	48	218
W	53	231	284
Total	223	279	502

- Specificity= 0.7623
- Sensitivity= 0.8279
- Precision= 0.8134
- Accuracy= 0.7988
- Error Rate= 0.2012

TABLE VII: CONFUSION MATRIX FOR MATCHES PLAYED BY INDIA TILL 2014 FOR 10 FOLD CROSS VALIDATION (SVM)

	L	W	Total
L	200	61	261
W	52	235	287
Total	252	296	548

- Specificity= 0.7936
- Sensitivity= 0.7939
- Precision= 0.8188
- Accuracy= 0.7937
- Error Rate= 0.2063

2) Naïve Bayes Classifier:

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. This algorithm has been applied to the datasets SL2014+ and IN2014+ as shown below to generate confusion matrix shown in Table VIII and Table IX.

DOI 10.17148/IJARCCE.2015.4642

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

TABLE VIII: CONFUSION MATRIX FOR MATCHES PLAYED BY SRI LANKA TILL 2014 FOR 10 FOLD CROSS VALIDATION (NAÏVE BAYES)

	L	W	Total
L	173	139	312
W	50	140	190
Total	223	279	502

- Specificity= 0.7758
- Sensitivity= 0.5018
- Precision= 0.7368
- Accuracy= 0.6235
- Error Rate= 0.3765

TABLE IX: CONFUSION MATRIX FOR MATCHES PLAYED BY INDIA TILL 2014 FOR 10 FOLD CROSS VALIDATION (NAÏVE BAYES)

	L	W	Total
L	210	178	388
W	42	118	160
Total	252	296	548

- Specificity= 0.833
- Sensitivity= 0.3986
- Precision= 0.7375
- Accuracy= 0.5985
- Error Rate= 0.4015

IV. RESULT

A. Accuracy

For comparison of the two algorithms, i.e. Support Vector Machine and Naïve Bayes, accuracy was used. Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Fig 1 and Fig 2 shows the comparison of SVM and Naïve Bayes (for all three data sets) for India and Sri Lanka respectively.

It can be inferred from these figures that the accuracy of models generated by the SVM have a much greater accuracy than the models generated by Naïve Bayes classifier.

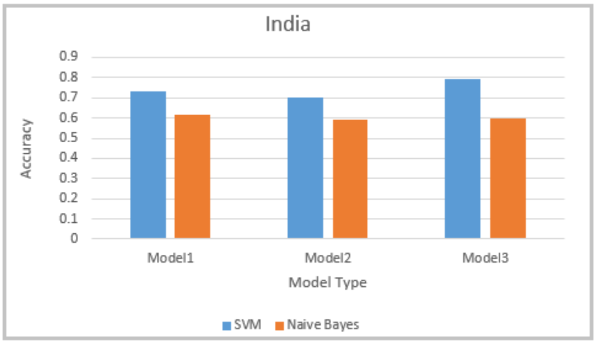

Fig 1. Comparison of SVM and Naïve Bayes for India

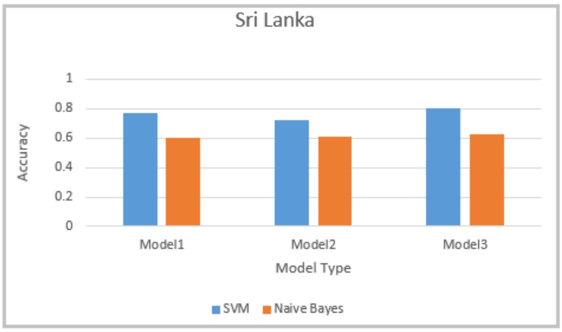

Fig 2. Comparison of SVM and Naïve Bayes for Sri Lanka

B. Model Outcome

The accuracy of a model checks out the combination of all the correctly predicted values by the total number of predictions made (TP and TN are considered). In combination they may show that the model is very good but when looked at separately do they still indicate the same. All the three factors- accuracy, type 1 error and type 2 error do not cover the cost of making a decision based on the predictions made by the model. These terms indicate that a true or false prediction made by model have the same cost. These true and false predictions made by the model are used to make certain decisions, and making a false decision based on the prediction machine made can be costly, even if the model may have a high accuracy rate.

In the world of betting the payoff can play a role and indicate how winning or losing and correct or incorrect predictions can make a person either lose or gain lots of money. This has been illustrated using the confusion matrix generated for Sri Lanka dataset (SR2007) by SVM. Let us say take a scenario that a bookie office gives you following bets:

- If you put a bet of Rs.1 on outcome that Sri Lanka would win.

Actual outcome: Sri Lanka wins you get 1.7 & if Sri Lanka loses you get zero.

- If you put a bet of Rs.1 on outcome that Sri Lanka would lose.

Actual outcome: Sri Lanka loses you get 2.2 & if Sri Lanka wins you get zero.

These are called the different cost payoffs. Effectively,

- If you put a bet of Rs.1 on Sri Lanka win: You get 0.7 if Sri Lanka wins and -1 if Sri Lanka loses.
- If you put a bet of Rs.1 on Sri Lanka Loss: You get 1.2 if Sri Lanka loses and -1 if Sri Lanka wins.

Table X shows the confusion matrix generated by Support Vector Machine on Dataset with matches played by Sri Lanka till 2007.

TABLE X : CONFUSION MATRIX FOR THE MATCHES PLAYED BY SRI LANKA TILL 2007 FOR 10 FOLD CROSS VALIDATION (SVM)

	L	W	Total
L	94	28	122
W	46	158	204
Total	140	186	326

DOI 10.17148/IJARCCE.2015.4642

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

Now suppose you use this model and would put one Rs. 1 bet on Sri Lanka win if models predicts win and you would Rs.1. On Sri Lanka loss if your model predicts Sri Lanka loss.

Probability (Model predicts Sri Lanka winning) = 204/326
Probability (Model predict Sri Lanka losing) = 122/326
Type 1 error = 46/204
Type 2 error = 28/122

Outcome of the Rs.1 bet on Sri Lanka winning given that model predicts that Sri Lanka would win is given by:
Outcome of win= (Payoff of wining)*(Probability (model correctly predicting win)) + (Bet Made)*(Probability (model incorrectly predicting win))
Outcome of win= 0.7 *(158/204) -1*(46/204) = 0.316667
Outcome of the Rs.1 bet on Sri Lanka losing given that model predicts that Sri Lanka would lose is given by:
Outcome of loss= (Payoff of losing)*(Probability (model correctly predicting loss)) + (Bet made)*(Probability (model incorrectly predicting loss))
Outcome of lose= 1.2*(94/122)-1*(28/122) = 0.695082.

If any of these values are negative, then you would rather not make a bet at all and thus outcome of win or loss becomes zero.

Expected Outcome of this model= Probability (model predicting win) * Outcome of win +
Probability (model predicting loss)*Outcome of loss
Expected Outcome of model =
204/326*0.316667+122/326*
0.695082= 0.458282.

The generalized formula used for Expected Outcome of model is:
Expected outcome={(TP+FP)/(P+N)} * { (TP/(TP+FP))* outcome(TP/(TP+FP))+ (FP/(TP+FP))* outcome (FP/(TP+FP))} +{(TN+FN)/(P+N)} * { (TN/(TN+FN))* outcome(TN/(TN+FN))+ (FN/(TN+FN))* outcome(FN/(TN+FN))}

Similarly the expected outcomes of different models was generated for SR2007 using Naive Bayes Classifier and for IN2007 using both SVM as well as Naive Bayes.
The generated results have been illustrated in the Fig 3 and Fig 4.

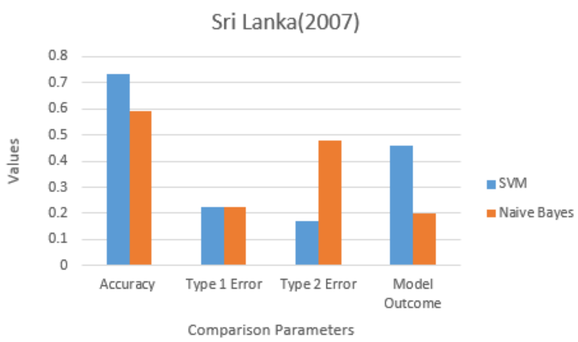

Fig 3. Comparison of SVM and Naïve Bayes on different parameters for Sri Lanka (2007)

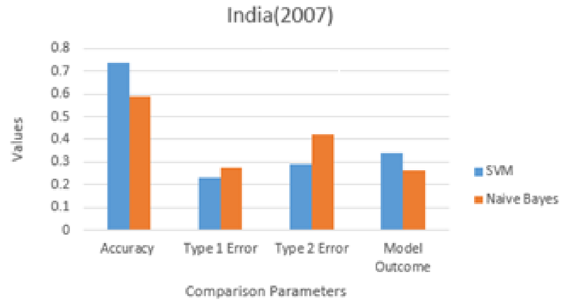

Fig 4. . Comparison of SVM and Naïve Bayes on different parameters for India (2007)

V. CONCLUSION

Till 2007, importance of "home field" advantage on One Day International cricket was statistically studied. This seems rational as playing at a home ground offers many benefits. Most of these are psychological in nature, such as familiarity with the playing grounds, the ability for participants to lodge in their homes rather than in a hotel, less likelihood of travel immediately prior to the game, and the support of the fans in attendance. However after addition of data from 2008 to 2014 the effect of home advantage have diminished in recent years. Possible reasons for this pattern are that there are young, enthusiastic players with international exposure, fans are willing to travel abroad to support their teams, emergence of new leagues in the recent years like Indian Premier League(IPL), Champions League T20, Big Bash League etc. which provide global exposure to all the players.
In addition, the strange result of the disadvantage of winning the coin toss for day time matches has also been observed. Even after winning the toss, the teams have lost a larger percentage of matches. This indicates that despite luck being on their side, they are prone to make bad decision. However, it was seen that winning the coin toss gives competitive advantage for "Day & Night" matches.
A model is said to be better if its accuracy and outcome values are higher and if their type 1 and type 2 errors are lower. It was found that SVM was proved to be a better model based on both the parameters used- accuracy and model outcome.

ACKNOWLEDGMENT

It gives us immense pleasure and satisfaction in submitting this paper on "Role of external factors on outcome of a One Day International (ODI) Cricket Match and Predictive Analysis". We would like to thank our college, **BMS College of Engineering**, for providing us with an opportunity to complete our final year project on this topic.
We would like to express our earnest thanks to our internal guide **Mr. Gururaja H.S.**, Assistant Professor, BMSCE who provided insight and expertise that greatly assisted the research.
Finally, we would also like to show our gratitude to **Janat Shah**, Director, IIM Udaipur, India and **Royal Denzil Sequeira**, Developer, Microsoft Research, India for sharing their pearls of wisdom and extra support and guidance throughout the course of this research.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

REFERENCES

[1] Michael J. Fry, Jeffrey W. Ohlmann(2015), Introduction to the Special Issue on Analytics in Sports, Part I: General Sports Applications in Interface. Publisher: Institute for Operations Research and the Management Sciences (INFORMS)

[2] M. J. Bailey & S. R. Clarke (2004), Market inefficiencies in player head to head betting on the 2003 cricket world cup. In Economics, Management and Optimization in Sport, S. Butenko, J. Gil-Lafuente & P. M. Pardalos, editors, Springer-Verlag, Heidelberg, pp. 185–202.

[3] M. J. Bailey & S. R. Clarke (2006), Predicting the match outcome in one day international cricket matches, while the match is in progress. Journal of Science and Sports Medicine, 5, 480–487.

[4]]D. Dyte (1998), Constructing a plausible test cricket simulation using available real world data. In Mathematics and Computers in Sport, N. de Mestre & K. Kumar, editors, Bond University, Queensland, Australia, pp. 153–159.

[5] Lindsey GR (1959), Statistical data useful for the operation of a baseball team. Oper. Res. 7(2):197–207.

[6] Lindsey GR (1961), The progress of the score during a baseball game. J. Amer. Statist. Assoc. 56(295):703–728.

[7] Lindsey GR (1963), An investigation of strategies in baseball. Oper. Res. 11(4):477–501

[8] James B (2001) The New Bill James Historical Baseball Abstract (Free Press, New York).

[9] Ananda Bandulasari(2007), Predicting the winner of a One Day International (ODI) Cricket match.