One Day International (ODI) Cricket Match Prediction in Logistic Analysis: India VS. Pakistan

Shanjida Chowdhury[1,*], K. M. Anwarul Islam[2,3], MD. Mahfujur Rahman[4], Tahsin Sharmila Raisa[5], Nurul Mohammad Zayed[6]

[1]Department of General Educational Development, Daffodil International University, Dhaka, Bangladesh
[2]Department of Business Administration, the Millennium University, Dhaka, Bangladesh
[3]Faculty of Business & Accountancy, University of Selangor, Selangor, Malaysia
[4]Department of Statistics, Comilla University, Kotbari, Cumilla-3506, Bangladesh
[5]Department of Business Administration, Daffodil International University, Dhaka, Bangladesh
[6]Department of Real Estate, Daffodil International University, Dhaka, Bangladesh

Abstract   Cricket is now a game of enchantment, refreshing and physical strength of 22 players. As the popularity of one-day international (ODI) games increases, it is essential to understand the game results' potential predictors. The advantage at home ground, coin-toss result, decision on first batting or fielding first, and day-to-day games are such popular cricket literature variables. The Indian subcontinent is like a game of thrill, war, friendship, and finally, and the best fighting teams are India and Pakistan. This study emphases a comprehensive analysis of the importance of these important predictors by various statistical calculations. For all ODI matches played by India & Pakistan from 1978 to 2019, information was manually collected from the website, www.espncricinfo.com. For purposes of model-building, logistic regression is applied retrospectively to data already obtained from previously played matches. Univariate, bivariate, binary and skewed logistic regression on the multivariate context are considered. In bivariate analysis, only the day/day-night format match is statistically significant at a 5% level of significance in favor of winning the Indian team. In binary logistic regression, the odds of winning team India was 70.6% times more for home ground and 2.28 times more for data compared to day-night times match. For skewed logistic, the odds of conquering for both teams increase, but the model performs comparatively worse than binary logistic regression.

Keywords   ODI, Cricket, Match, Prediction, India, Pakistan, Binary Logistic Analysis

1. Introduction

"ODI or One Day International "was a cricket form in the 1960s as the five-day-long test cricket alternative. Before one day match, it was a five-day test match with both teams wearing white outfits. Every team has to bat and bowl twice, depending on the score. It was a long process, and most of the matches end up with a draw to solve this problem and add variety to cricket. "One-day International cricket" (ODI) was introduced. A One-day international match is a 50 over match with both teams wearing colorful jerseys. Both team bats and bowl and scores, the team who scores most will win the game, and there is no scope of a draw because if the scores are level, then each team will bat and bowl again for one over which is called "Super over." Talking about India and Pakistan's teams was the most controversial and rival teams in cricket history because of the tense relationship between

them.

These two teams had faced each other 132 times, so by just this information, it's tough to predict which team will win, but by some statistical analysis, we may come to an absolute conclusion and may be able to predict which team is going to win.

This is the cricket scenario that is closely related to an individual's real-life condition or culture as a whole that has proven that fate prevails above everything others. The game is cutthroat in real life, and both rivals are as strong and effective with their unique skill set as the world's mighty cricket team. Competitors also put in much hard work but struggle to achieve the desired outcomes at crucial moments like the two former winners, India and Pakistan. Team strategists currently rely on a combination of personal experience and team-building. Inherently human experts' methodology is to extract and leverage important information from past and current game statistics. It can also be helped to make strategic decisions to increase winning chances. This research aims to study predicting the game results before the game starts, based on the statistics and data available from the data set.

2. Literature Review

Cricket is a sport where data mining can be applied in several different ways. For example, in a one-day international (ODI) format, an infinite range of inquiries can be answered with aid from data analytics. Previous experiments have shown that specific probabilistic associations in match outcomes can be proof of variables like a home-field benefit, coin-toss result, and batting first or second. (De Silva and Swartz [1], Allsopp [2], and Bandulasiri [3]). Elderton and Wood [4] fit the geometric distribution to individual scores based on results from test cricket in the pre-computer days. Kimber & Hansford reject geometric distribution and gain probabilities in test cricket using product-limit estimators for selected sets of individual scores. Their examination depends on showing traditional distribution runs [5]. Some of the work seen here by Dyte [6] simulates batting results between a defined test batsman and bowler using career batting and bowling averages as main inputs regardless of match status.

Various researchers used various prediction models. Ganeshapillai and Guttag [7] established a prediction model for baseball that determines when, as the game progresses, to change the starting pitcher. It is very much related to our work-flow, where the amalgamation of previous data and game data was used to forecast the success of a pitcher. The One Day International (ODI) is the most frequent and famous form of cricket, with over 50-overs per side being played. Winning is the ultimate objective, as it's common in sports games. Some research evaluates the extent of the success, De Silva [8], but most consider the factors that influence winning. Kaluarachchi

and et al. (2010) take into account various factors that influence the game, including the benefit of the home side, the day/ night impact and toss, etc., and use the Bayesian classifier to predict the match's outcome. Sankaranarayanan et al. [9] used the method of machine learning to predict the result of a one-day match based on previous experience and game experience. Sohail Akhtar and Philip Scarf [10] have predictions, session by session, matching results in test cricket in play. Match outcome probabilities are forecast using a series of multinomial logistic regression models at the start of each session. These probabilities will make it easier for a team captain or management to consider the coming session with an offensive or defensive batting strategy. Madan Gopal Jhawar and Vikram Pudi [11] predict the consequence of ODI cricket matches with an approach focused on team composition. Shah [12] developed a model forecasting match results for each ball played. The result of the Duckworth-Lewis formula match for live matches will be expected. The probability is determined for each ball bowled, and the probability figure is plotted. In order to resolve the home-field advantage in ODI games, Fernando et al. applied a logistic regression technique [13].

Bailey & Clarke [14] focused on how external factors decide ODI cricket matches. Some of the more influential factors include the home ground advantage, team quality (class), and current form. Other work is undertaken to forecast match outcomes in ODI cricket based on external factors: Bandulasiri[3], using the Logistic Regression methodology to investigate the predictive importance of different features and construct a model. In this paper, we expand his work done by using Logistic Regression and its odds ratio to measure match outcome probabilities. This study works for finding out the impact of winning matches of the Indian team through toss winning, home ground favor, day or day-night play, and decision on first batting. This study works for finding out the impact of winning matches of both teams through toss winning, home ground favor, day or day-night play, and decision on first batting. In the history of cricket, these two teams are facing on the ground and clog with geographical lines. Last 2 or 3 decades, the matches are not seen as much as before due to political, geographical, and historical enmities. This study enlightens to revive their previous prestige and emphasize more action to come in friendship and arrange more series on paradises and root out all distances. Also, to predict ODI Cricket Match, this study considered India and Pakistan cricket teams only because of exemplary purposes. The study result will apply to any other country.

3. Materials and Methods

Data were manually obtained from a website, www.espncricinfo.com, for all ODI matches between 1978 and 2019 played by India & Pakistan. The data

gathered were subjected to a cleanup process where specific matches were omitted from the study due to particular factors such as excess lousy weather or where one group was much better than the other (ranked team playing non-ranked teams). (ranked team playing non-ranked teams). The analysis also deleted tied games. Therefore, we just review games containing clear decisions. Save the data collection in a comma-separated format. Our entire methodology is encircled through univariate, bivariate & Multivariate analysis. Statistical Software Used in this study is SPSS and Stata 15.0 version to determine the result of Skewed logistic regression.

Multivariate analysis: On multivariate analysis, our goal is to examine one variable's influences on other or other sets of variables. Moreover, we try to know whether the variables are involving importance or not.

Logistic Regression Model:

A brief description of the model is given below. Suppose $p = Pr(y = 1)$ is the probability that an individual has diagnosed asthma, and thus $(1 - p)$ represents the probability of not diagnosed. Now, p_i is non-linearly related to the explanatory variables as

$$p_i = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$

Where p_i ranges between 0 to1 and X represent explanatory variables, and β refers to the regression coefficients. The odds ratio is

$$\frac{p_i}{1 - p_i} = \exp(X\beta)$$

Taking natural logarithm into both sides of the equation would provide the equation for the logistic regression technique, and which is as follows:

$$ln\left(\frac{p_i}{1 - p_i}\right) = X\beta$$

Here, $ln\left(\frac{p_i}{1-p_i}\right)$ is the log odds ratio. Skewed logistic regression has also been employed in this study.

4. Results and Discussion

Univariate analysis:

In cricket, the two most battleships anchor to each other are India & Pakistan. India, two names are enough for making goosebumps in a cricket fan. In India's context, it's a name where Kapil dev, Sunil Gavaskar, Sachin Tendulkar, Sourav Ganguly, Rahul Dravid & notable match winners are still remembered & respected. On the other hand, the "Unpredictable team" is also famous for Imran Khan, Abbas, Sayeed Anwar, Wasim Akram,

Waqar Yunus, and other playmakers are still acknowledged. During 50 years, 128 matches come in one favor, where Pakistan won 73 times and India on 55.

In playing cricket, the home atmosphere is a plus point for taking fewer challenges and more success. Out of 128 matches, team India played slightly more than the Pakistan team. Whether a team will lead or chase run is decide upon the basis of toss winning. Here from table 1, it can be seen that about 51 percent of tosses were own by team India. First bating is an asset if there is enough experienced batsman available and an opening opportunity to play a freehand stroke. However, to adjust to the condition or playground, chasing is better. 53.9% time Indian team decided to bat on second innings. On the eve of the 21st century, ICC recognized match playing on the night in a test match. It's still a good practice for ODI format to play matches in the Day-night form.

Table 1. Descriptive statistics of India & Pakistan cricket match

Characteristics	Category	Frequency	Percent
Time frame	1978-1990	32	25.0
	1991-2000	49	38.3
	2001-2016	47	36.7
Venue	Host India	30	20.9
	Host Pakistan	25	23.5
	Others	73	19.4
Toss	India Win	65	50.8
	Pakistan Win	63	49.2
Batting/ chasing	Batting first	59	46.1
	Chasing	69	53.9
Day/Day night	Day	86	67.2
	Day-night	42	32.8
Total		128	100

Bivariate analysis:

In this study, winning games according to both teams is considered dependent, and others are the independent variable. Table 02 shows the result of Cross-tabulation of variables with wining chance of Team India

Venue vs. winning match: Out of 50 winning matches, 10 (20%) matches are on home ground while 32 (64%) are played on other venues. On the other hand, 46 (54.8%) matches are lost on the outdoor ground.

Day-night/day matches vs. winning match: out of 128 matches, 42(32.8%) matches are played on the day-night format. In this format, the chance of winning matches of the Indian team is better, 20 matches won out of 42 played. On the other hand, the proportion of losing games is higher for daytime matches played.

Toss winning vs. winning match: The decision to make a better score or pressure the opposite team through a great bowling line up largely depends upon winning thetoss. Team India won 65 times toss, of which 63

matches lost.

Batting first vs. winning match: In the first batting of team India, 26 times won out of 59 matches played. However, at chasing, the winning strike is less than half of losing matches.

Our logistic regression is assessed through the dependent variable of the wining of both teams. Here table 03 represents the result of **multiple logistic regression for winning for team India and Pakistan**. On reference category of Pakistan winning, the only day-night factor is significant & it assures winning of 134.3 percent if a match is playing on a day-night atmosphere. If playing a match in the home is 1.46 (0.61-3.55) times greater possibility than to play away matches. We test another logistic regression known as skewed logistic regression (scobit) (under a robust standard error to ensure a more defined model) for further scrutiny. Under the ML estimator, unlike logit, the regressors' consequences on the likelihood of success are not restricted to be larger when

the potential is 0.5. Exploratory factors can fairly have their most significant effect when the chance of success is 0.3 or 0.6. This situation is related to this study. Under this added tractability, outcomes attributable to scobit function (not logit function) could be distorted and the 0.5 chance of success cannot be forced to be mirror-symmetric. Logit functions stronger (AIC=185.45) from these two models that it is distorted (AIC=187.45) because parameters are not limited. On the other hand, using the reference category of India win, the chance of winning a match is 0.74 times significantly more in a day as compared to day-night matches. Others don't have related considerably to winning chance of Pakistan team, though the ground on both India as well as own ground is favorable for Pakistan. Under model selection, logistic is better as before. Adjusted R-square is comparatively low (0.082 for binary, 0.068 for skewed logistic). Due to the lack of more influential variables, we assess these two types of regression under binary outcomes.

Table 2. Cross-tabulation of variables with wining chance of Team India

			Result in favor of Team India		Total	Chi-square	df	Cramer's- V
			Win	Loss				
Venue	India host	N	11	19	30	0.646	2	0.07
		%	8.6%	14.8%	23.40%			
	Pakistan host	N	11	14	25			
		%	8.6%	10.9%	19.5%			
	Others	N	33	40	73			
		%	25.8%	31.3%	57%			
Day /day-night	Day	N	35	51	86	7.911	1	0.78
		%	27.3%	39.8%	69.40%			
	Day-night	N	20	22	42			
		%	15.6%	17.2%	32.80%			
Toss	win	N	27	38	65	0.74	1	0.02
		%	21.1%	29.7%	50.8%			
	loss	N	28	35	63			
		%	21.9%	25.3%	49.20%			
Batting first	1st	N	26	33	59	1.25	1	0.02
		%	20.30%	25.6%	46.10%			
	2nd	N	29	40	69			
		%	22.70%	31.30%	53.90%			
	Total	N	**55**	**73**	**128**			
		%	**43.0%**	**57.00%**	**100.00%**			

Table 3. Multiple logistic regression of wining chance of team India (reference category Pakistan)

	Reference category Pakistan						Reference category India					
	Binary logistic			Skewed logistic			Binary logistic			Skewed logistic		
	Adj. OR	Lower Adj. OR.	Adj. OR	Lower Adj. OR.	Upper Adj.OR.	Adj. OR	Adj. OR	Lower Adj. OR.	Adj. OR	Lower Adj. OR.	Upper Adj.OR.	Adj. OR
Intercept***	0.99			1.789			0.99			1.91		
Toss												
India won	1.102	0.543	2.238	1.158	0.169	7.926	0.908	0.447	1.843	0.916	0.350	2.399
Pakistan won (RC)												
Batting/ Chasing												
Batting first	0.879	0.431	1.792	0.824	0.078	8.740	1.138	0.558	2.32	1.124	0.352	3.583
Chasing												
Day or day-night***												
Day	1.343	0.636	2.837	1.516	0.024	94.978	0.744	0.352	1.572	0.770	0.050	11.889
Day night (RC)												
Venue**												
Home India	1.465	0.605	3.545	1.554	0.030	79.307	0.683	0.282	1.652	0.751	0.023	25.071
Home Pakistan	1.053	0.42	2.64	1.688	0.014	206.34	0.949	0.379	2.38	0.716	0.017	30.859
Others (RC)												

5. Conclusions

Cricket is now a populous game in the South Asian Sub-continent. India and Pakistan are the most competitive two rivals of this 50 overs format. This study enables to finding hindsight of wining match towards team India. Toss winning, venue, day or day-night format, and batting at first are considered independent variables towards the Indian winning strike. For measuring association, Cross tabulation along with Chi-square statistic is observed. Also, for multivariate analysis, binary logistic regression is performed. Day-night format and venue give a statistically significant result in winning the chance of Team India. For better checking out, we check robustness under Skewed logistic regression. It enables us to perform ML estimator for skewed cases of probability within the dichotomous variable. Though the chances of achieving a win increase under this robust estimation, model information criteria and adjusted R square has performed poorly. Interestingly, De Silva and Swartz [1] addressed the home-field advantage in ODI cricket matches and found it a significant advantage for the home team. This study's limitations are using a dummy as some crucial player's involvement or century of a batsman, or even five wickets haul is essential for winning a 50-50 overs match. These significant dummies are unobserved due to a lack of data available from the official website. Also, providing an excellent total score (suppose, 250 or more is a fighting score) is not taken in this study. Still, this study entails an overview of historical data of playing matches between the ultimate two rivals. Like cricket, life is unpredictable, and one needs a mixture of committed effort and fate, a helping hand to produce the desired outcome.

6. Future Work

The statistical analysis was carried out in this study to forecast match outcomes, suggesting some avenues for future research work. Attributes such as temperature, humidity, event, and wind speed could not be used in analysis due to their large number of missing values. It is possible to further develop the proposed model by taking specific attributes into account. In addition, for test and Twenty20 game formats, a comparable analysis of predicting match outcomes may also be carried out.

REFERENCES

[1] De Silva, B.M., Swartz, T.B., "Winning the coin toss and the home team advantage in one-day international cricket matches," Department of Statistics and Operations Research, Royal Melbourne Institute of Technology; 1998 Mar.

[2] Allsopp, P., "Measuring team performance and modelling the home advantage effect in cricket," Doctoral dissertation, Swinburne University of Technology, 2005.

[3] Bandulasiri, A., "Predicting the winner in one day international cricket," Journal of Mathematical Sciences & Mathematics Education, 3(1), 6-17, 2008.

[4] Elderton, W., & Wood, G. H., "Cricket scores and geometrical progression," Journal of the Royal Statistical Society, 108(1/2), 12-40, 1945.

[5] Kimber, A. C., & Hansford, A., "A Statistical Analysis of Batting in Cricket," Journal of the Royal Statistical Society, 156(3), pp.443-455, 1993.

[6] Dyte, "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, 1998.

[7] Ganeshapillai G., Guttag, J., "A data-driven method for in game decision making in MLB: When to pull a starting pitcher", In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 973–979, 2013.

[8] De Silva, B. M., Swartz, T. B., "Estimation of the magnitude of the victory in one-day cricket," Australian & New Zealand Journal of Statistics, vol. 43, pp. 1369-1373, 2001.

[9] Sankaranarayanan, V. V., Sattar, J., Lakshmanan, L.V.S., "Autoplay - A data mining approach to ODI cricket simulation and prediction," In: Proceedings of the 2014 SIAM International Conference on Data Mining, pp. 1064–1072, 2014.

[10] Akhtar, S., Scarf, P., "Forecasting test cricket match outcomes in play," International Journal of Forecasting, vol. 28, no. 3, pp. 632-43, 2012 Jul 1.

[11] Jhanwar, M.G., Pudi, V., "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach'" InMLSA@ PKDD/ECML, 2016 Sep 19.

[12] Shah, P., "Predicting Outcome of Live Cricket Match Using Duckworth-Lewis Par Score," International Journal of Systems Science and Applied Mathematics, vol. 2, no. 5, pp.83-86, 2017, DOI: 10.11648/j.ijssam.20170205.11.

[13] Fernando, M., Manage, A., Scariano, S., "Is the home-field advantage in limited overs one-day international cricket only for day matches?," South African Statistical Journal ,47, pp. 1-13, 2013.

[14] Bailey, M., & Clarke, S. R., "Predicting the match outcome in one day international cricket matches, while the game is in progress," Journal of sports science & medicine, 5(4), pp. 480, 2006.