Analytics of Batting First in Indian Premier League Twenty20 Cricket Matches

Christopher R. Brydges

NIH West Coast Metabolomics Center

University of California, Davis

christopherbrydges@gmail.com

Manuscript DOI: 10.31236/osf.io/jq564

Please cite as:

Brydges, C. R. (2021). *Analytics of batting first Indian Premier League twenty20 cricket matches*. SportRxiv. https://doi.org/10.31236/osf.io/jq564

Purpose: Use ball-by-ball data from the Indian Premier League cricket tournament and machine learning techniques to predict match outcomes based on events occurring in the first inning of a match.

Approach: Twelve predictor variables were entered into machine learning models (forward stepwise logistic regression using Akaike's Information Criterion (AIC); forward stepwise logistic regression using Bayesian Information Criterion (BIC); random forests; naïve Bayes classifier), with match outcome as the dependent variable.

Findings: The AIC model reported the highest accuracy in both the training and test datasets (69.92% and 67.18%, respectively). This model contains total runs scored, winning the coin toss, and home-ground advantage as positive predictors, and number of balls with no runs scored and number of balls with one run scored as negative predictors. All four models found that total runs scored in an inning was the most important predictor of match outcome, and no model included number of wickets lost as a predictor, although there could be an indirect effect through total runs scored.

Originality: This study is novel in that it used both pre-match variables (home-ground) advantage and real-time measures (e.g., how many runs were scored in the powerplay) in a machine learning context to classify match results. The results can be used to adapt in-game tactics to maximize advantages of batsmen in favorable contexts.

Keywords: Cricket, Data science, Machine learning, Prediction, Sport analytics

Article classification: Research Article

Analytics of Batting First in Indian Premier League Twenty20 Cricket Matches

Cricket is estimated to be the second-most popular sport in the world, behind only soccer (1). The sport is especially popular in India, which hosts the annual Indian Premier League (IPL): A competition of the shortest form of the game of 20 overs per side, in which the world's best players play for franchises based in various Indian cities. In 2017 - only the eleventh season of the competition - the IPL was valued at $5.3 billion (2). Additionally, sponsorships in 2017 totaled $1 billion, which was more than the $892 million made in sponsorships in Major League Baseball in the same year (2). As such, being able to predict performance can potentially have major financial implications for players and coaches (whose salaries are dependent upon performance), as well as teams (who may command greater sponsorship revenues if the team is successful). In-game tactics can be developed and applied by coaches and analysts based on events occurring in previous matches and/or data related to individual players (3).

The purpose of this paper was to use ball-by-ball data from the IPL to investigate four research questions regarding accurate prediction of match outcome: First, what is a winning score when batting first? Second, to what extent is the probability of victory affected by the loss of wickets in the powerplay (the first six overs)? Third, is there a particular phase of play (i.e., the powerplay, lull, ramp-up, or death overs) that is especially predictive of match outcome? And fourth, teams that bat first in each game were used as cases to see if any one out of twelve variables (e.g., number of runs scored, number of wickets lost, etc.) are predictive of match outcome.

Method

Data Source

Ball-by-ball and match result data were taken from Kaggle (4). All analyses were conducted in R 4.0.4 (5), and a script to download the data and reproduce all analyses and figures is publicly available from https://osf.io/ze5qp/.

Pre-Processing

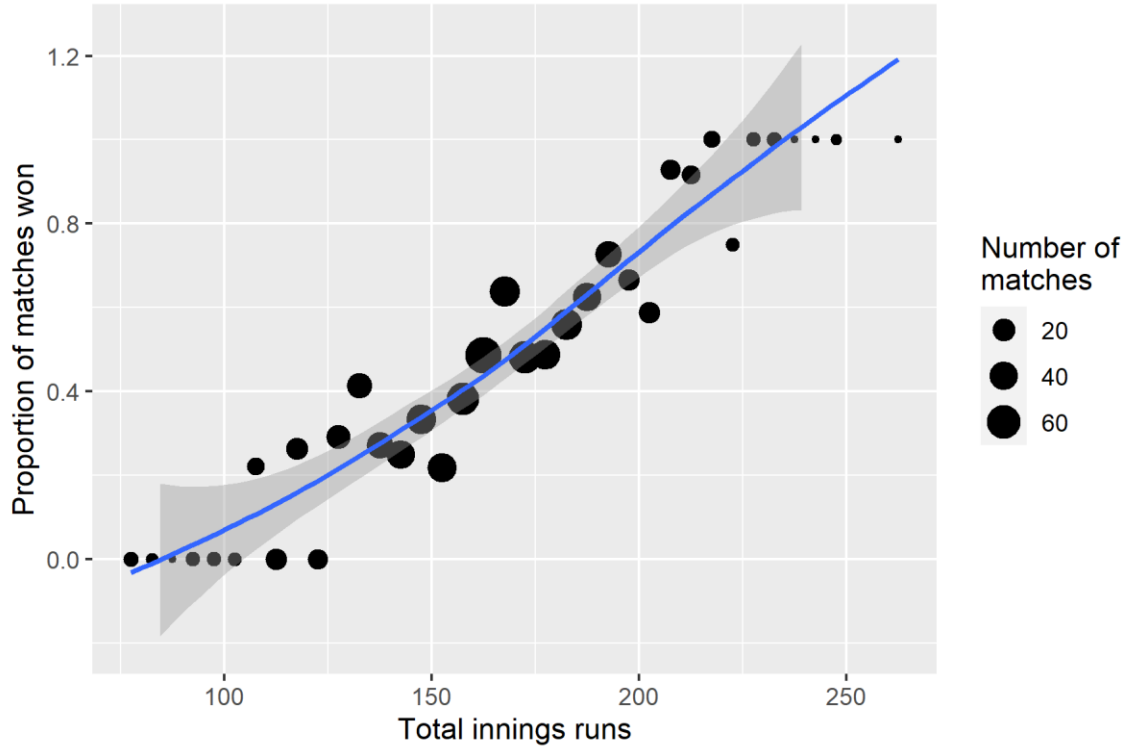
All regular season and finals matches from the inaugural 2008 season up to the recently completed 2020 season were entered into pre-processing. For all analyses, matches that were ties or abandoned were removed (i.e., only wins and losses were used), and any match that was shortened (i.e., due to inclement weather conditions) was also excluded. This resulted in a final sample size of 780 matches.

Analyses and Results

Analysis 1: What is a winning score when batting first?

Figure 1 displays the association between proportion of matches won and number of runs scored by the team batting first. Note that due to the continuous nature and large range of total runs scored in an entire innings (median = 163, minimum = 67, maximum = 263, interquartile range = 39), runs were binned into five run groups (e.g., < 80, 80-84, 85-89, etc.), and the proportion of matches that resulted in a victory was calculated for each bin for this figure.

Figure 1. Bubble plot showing the proportion of matches won and runs scored. Point size represents the number of matches within the five-run bin. The blue line is the loess regression line of best fit, and the shaded region around it is the 95% confidence interval.

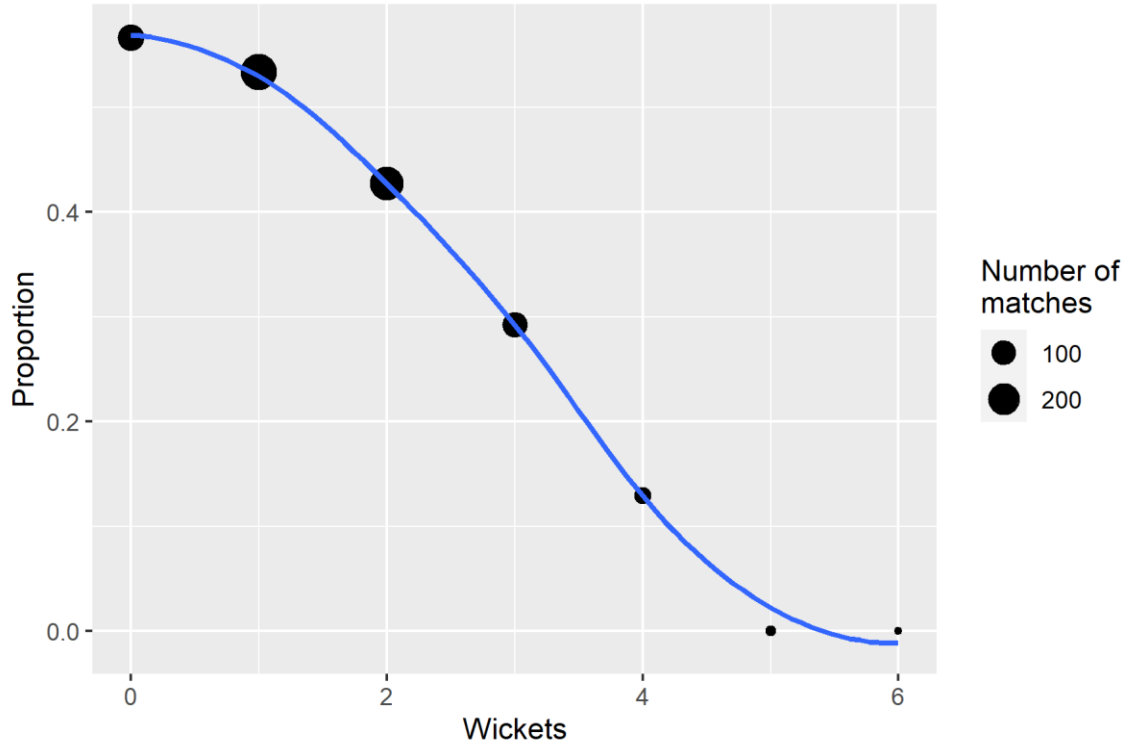


Clearly, there is a very close association between scoring more runs and the probability of victory. A simple logistic regression was conducted to analyze this. Unsurprisingly, total runs scored was predictive of match result ($Z = 10.45$, $p < 2.2\times10^{-16}$). Based on this analysis, a team has a 50% chance of winning if they score 170 runs in their inning, and an 80% chance of winning if they score 210. That is, all other things being equal, a 'par score' when batting first in the IPL is 170. As of March 7th 2021, only seven IPL matches have ever seen a team batting second score at least 210, and only four of these resulted in victories (6).

Analysis 2: How is the probability of victory affected by loss of wickets in the powerplay?

This analysis was inspired by Alex Wakely, then captain of English county Northamptonshire, who stated in 2016 that teams that lose three wickets during the powerplay end up losing 83% of T20 matches (7). Figure 2 shows the proportion of matches won by the side batting first, given the number of wickets lost during the powerplay.

Figure 2. Bubble plot showing the proportion of matches won and wickets lost in the powerplay. Point size represents the number of matches where the corresponding number of wickets have been lost. The blue line is the loess regression line of best fit.



Visual inspection of Figure 2 suggests that teams batting first can give themselves a slight advantage if they lose either zero wickets or one wicket during the powerplay, and that the disadvantage of losing wickets increases as more wickets are lost. A logistic regression confirms this (Z = -6.314, p = 2.72×10^{-10}). Table I shows the probabilities of victory for the team batting first, given the number of wickets they lose during the powerplay.

Table I. Probability of victory for team batting first based on number of wickets lost during the powerplay.

Wickets	Probability of winning the match
0	0.627
1	0.515
2	0.402
3	0.298
4	0.212
5	0.145
6	0.097

With regards to Wakely's claim (7), discrepancies between his number (17% of matches where a team loses three wickets in the powerplay result in victories) could be explained by a number of

potential factors: First, it is likely he is speaking about English T20 matches, which could potentially differ from IPL matches. Second, this value could have changed since he stated it. Third, this analysis only examines the first innings of each match. Additionally, if weighted proportions are used, losing three *or more* wickets in the powerplay results in a probability of victory of 0.263, which is slightly closer to his estimate (though still quite a way off, and certainly still bad for the batting team).

Analysis 3: Is there a phase of play that is predictive of match outcome?

Four phases of an innings were defined following Kimber's (8) suggestions: Powerplay (the first six overs where there are fielding placement restrictions), the lull (overs 7-12), the ramp-up (overs 13-16), and the death (overs 17-20). Mean run rates for each over are shown in Figure 3, and boxplots showing run-rates and pairwise comparisons for each phase are displayed in Figure 4. In both figures it is apparent that runs are scored most slowly during the lull period immediately following the powerplay, and that runs are scored fastest in the death phase at the end of the inning. A logistic regression, using auto-scaled runs scored in each of the four phases as predictors of match result, found that runs scored in all four phases are significantly predictive of outcome (powerplay $Z = 4.146$, $p = 3.39\times10^{-5}$; lull $Z = 4.402$, $p = 1.07\times10^{-5}$; ramp-up $Z = 3.770$, $p = 1.63\times10^{-4}$; death $Z = 5.912$, $p = 3.38\times10^{-9}$)

Figure 3. Mean run rates for each over of the first innings of IPL matches.

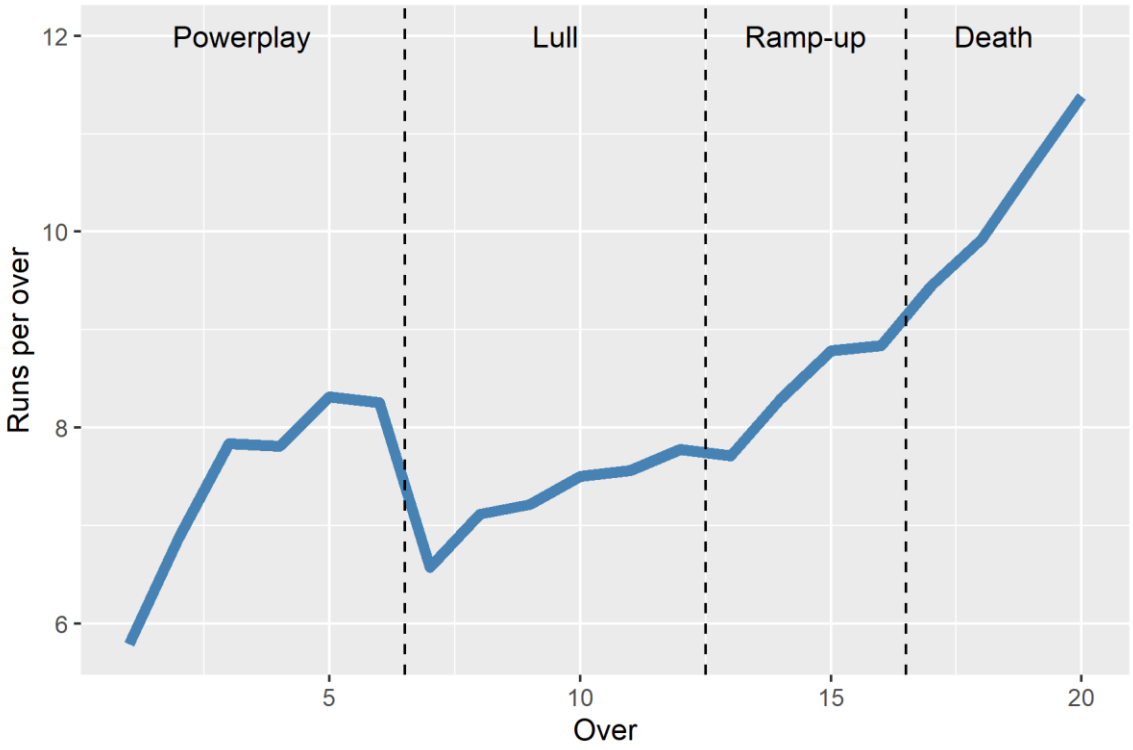
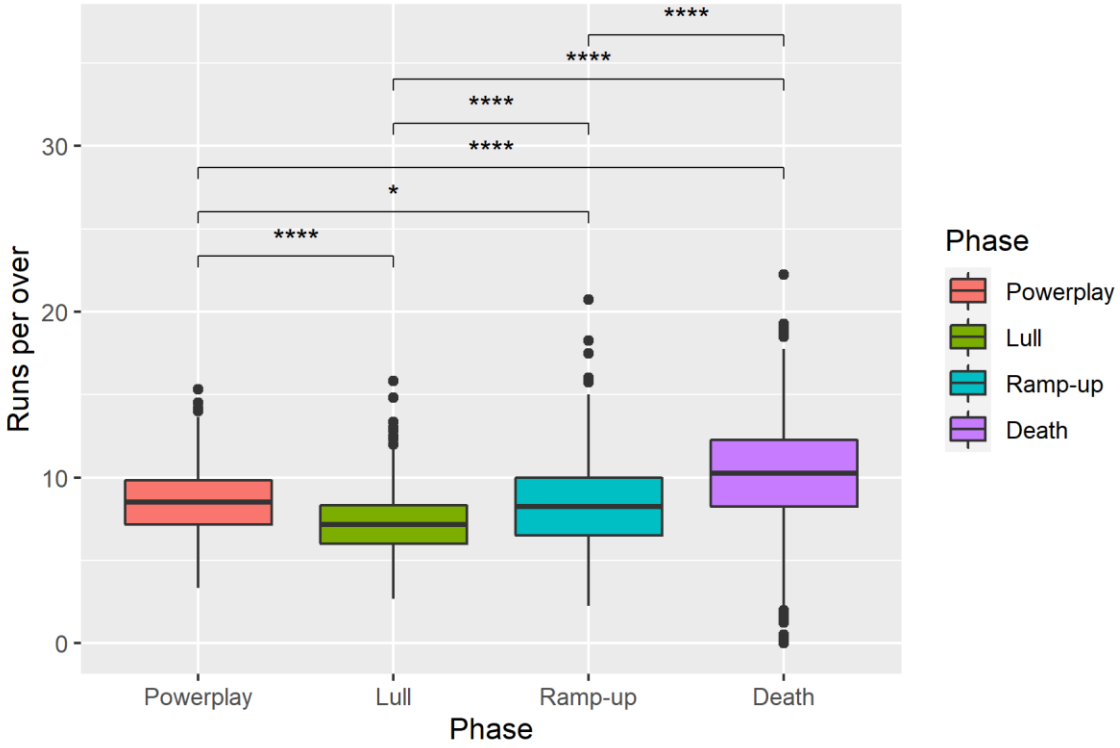


Figure 4. Boxplots of run rates for each phase of the first innings of IPL matches. Asterisks denote statistical significance for paired-samples *t*-tests between for each pairwise comparison. * = $p < 0.05$, **** = $p < 0.0001$.



Analysis 4: What factors in the first innings predict match outcome?

Some previous research has performed machine learning analyses on cricket data. Recently, Kapadia et al. (9) investigated the potential effects of home-ground advantage and winning the coin toss (i.e., having the advantage of being able to decide to bat or bowl first in the match), and reported prediction accuracy of match outcome below 70%, though the researchers acknowledged that pre-game conditions (i.e., home-ground advantage and winning the coin toss) may not be as important as live, ball-by-ball data.

The current study included the following twelve variables as predictors of match outcome: Total runs scored in the inning; Number of runs scored in the powerplay; Number of balls that had no runs scored from them; Number of balls that had one run scored from them; Number of balls that had two runs scored from them; Number of balls that had three runs scored from them; Number of balls that had four runs scored from them; Number of balls that had six runs scored from them; Number of wickets lost in the inning; Number of wickets lost in the powerplay; Winning or losing the coin toss; Home-ground advantage. The first ten variables, which are all continuous, were auto-scaled (i.e., centered to zero, with a standard deviation of 1) before being entered into analyses. The data were randomly split into training (75%) and test (25%) sets to validate the models.

In the training dataset, it was found that teams batting first only won 45.26% (358/780) of matches, which implies that there is already a statistically significant advantage to batting second ($\chi^2(1) = 6.832$, $p = 0.009$). Initial exploratory modeling was conducted, with single logistic regressions conducted separately for each predictor variable, with match result as the dependent variable. With the exception of the number of balls that had one run scored from them, number of balls that had three runs scored from them (probably due to this being a rare occurrence), and whether the toss was won, all predictor variables were significantly predictive of match result ($p < 0.05$).

Four models were tested: A forward stepwise logistic regression based on Akaike's Information Criterion (AIC); a forward stepwise logistic regression based on the Bayesian Information Criterion (BIC); Random Forests; and Naïve Bayes classifier. These models are referred to as the AIC, BIC, RF, and NB models hereon.

The AIC model included total runs scored in the inning, winning the coin toss, and home ground advantage as positive predictors (i.e., greater runs scores, winning the toss, and/or playing at home increased the probability of the team winning), and number of balls that had no runs scored from them and number of balls that had one run scored from them as negative predictors (i.e., a greater number of balls with zero or one run scores from them decreased the probability of the team winning). This last finding - where a run is still being scored - might be initially counterintuitive, but given that the average total runs scored by a winning team batting first is 176, which corresponds to 1.47 runs per ball on average, a team needs to minimize the number of balls where a low number (i.e., zero or one) of runs are scored from them. Table II shows the regression estimates of these predictors. The AIC of this final model was 689.87.

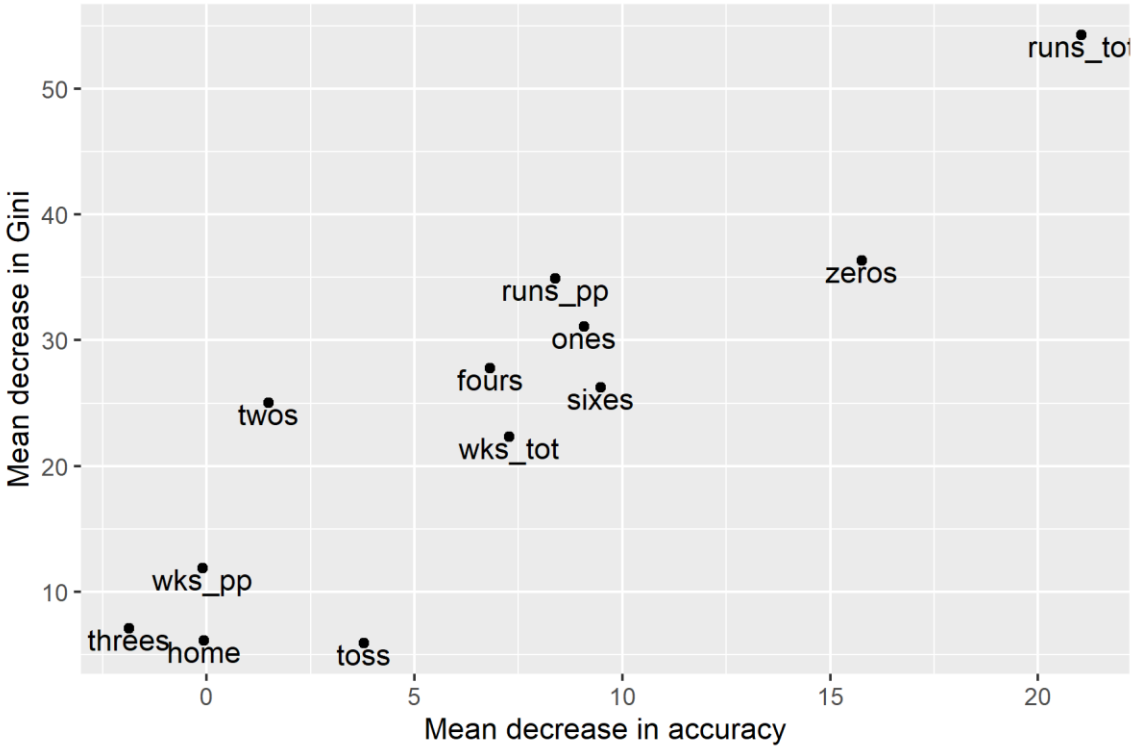
Table II. Regression estimates of the AIC model.

Predictor	Estimate	Standard Error	*p*
Intercept	-0.555	0.160	0.0005
Total runs	0.740	0.164	6.54×10^{-6}
Coin toss won	0.376	0.192	0.0503
Home-ground advantage	0.272	0.187	0.1462
Number of zeros	-0.572	0.201	0.0044
Number of ones	-0.350	0.151	0.0201

The BIC model only included total runs scored in the inning as a predictor (more runs was associated with greater probability of winning). The AIC of this model was 696.46.

The RF model identified total runs, number of balls with no runs scored, then number of balls with one run scored as the most important factors (same as the AIC step-wise model), followed closely by number of balls which has six runs scored of them. In contrast to the AIC model, winning the coin toss and home-ground advantage were found to not be important predictor variables. Figure 5 shows predictor variable important, as measured by decrease in mean accuracy and Gini.

Figure 5. Scatter plot of mean decrease in accuracy and Gini for each predictor variable in the RF model.



Lastly, the NB model reported that total runs was by far the most important predictor variable, followed by number of balls with no runs scored, number of balls with six runs scored, total number of wickets lost, and number of balls with four runs scored (see Figure 6). In this model, the number of balls with one run scored, winning the coin toss, and home-ground advantage were all relatively unimportant.

Table III displays the accuracy and confusion matrices of the four models. All models had similar accuracy, although the AIC model did perform slightly better than the others. The NB model had a noticeable tendency to classify results as wins, compared to the other three models.

Table III. Accuracy and confusion matrices of the four models using the training data.

		Predicted Result	
Model (% correct)	Actual result	Win	Loss
AIC (69.92%)	Win	169	97
	Loss	79	240
BIC (67.86%)	Win	166	100
	Loss	88	231
RF (68.55%)	Win	161	105
	Loss	79	240
NB (69.74%)	Win	191	75
	Loss	102	217

Lastly, these models were applied to the test data. It was found that the AIC model had the highest accuracy (67.18%), followed by the NB model (65.64%), the RF model (62.56%), and the BIC model (62.05%). Therefore, it was concluded that the AIC model that included total runs scored, number of balls with zero runs scored, number of balls with one run scored, winning the coin toss, and home-ground advantage, was the best model for this data. Confusion matrices for the classification of the test data are presented in Table IV. As with the training data, the NB model had a greater tendency to predict wins than the other three models.

Table IV. Confusion matrices of the four models using the test data.

		Predicted Result	
Model	Actual result	Win	Loss
AIC	Win	47	40
	Loss	24	84
BIC	Win	40	47
	Loss	27	81
RF	Win	44	43
	Loss	27	81
NB	Win	50	37
	Loss	30	78

Discussion

The current study aimed to use ball-by-ball data from the IPL to investigate four research questions regarding accurate prediction of match outcome. Overall, the four machine learning models that were applied to the test data set were all quite similar in their accuracy, but the forward stepwise logistic regression using AIC model was the most accurate and was relatively simple, as it only contained five predictor variables. It is intuitive that all models included total runs scored as the most important predictor variable: ultimately, whichever team scores the most runs wins the match. It is also notable that the number of balls that had four runs scored from them was not important in any model, as scoring four runs from a ball is generally considered as successful for the batsman, and is relatively easier to achieve than scoring six runs (nb. it is possible to score five runs from one ball, but this is highly unusual). However, given that number of balls that zero runs and one run were scored from were both negatively predictive of match outcome, a logical extension of this is that the number of balls that runs were scored from is positively predictive, with the caveat that more than one run per ball is scored. Seeing as both zero runs and one run are harmful to a team's chances of winning, it may also be the case in the future that a batsman refuses a run from a ball so he can face the same bowler again on the next ball, if the bowler is known to bowl to the batsman's strengths (i.e., it is likely the batsman will be able to score four or six runs from the following ball).

Additionally, the number of wickets lost in the first innings does not appear to have any impact on match outcome. While it is likely that number of wickets lost may indirectly affect match outcome through runs scored (there is a strong negative correlation between number of wickets lost and runs scored, $r = -0.59$, $p < 2.2 \times 10^{-16}$), it is also likely that to some degree, teams batting first are

not optimally using their batting resources. Specifically, teams may be batting too defensively in order to conserve wickets, when they can attack more (10), especially in the lull overs immediately following the powerplay (see Figures 3 and 4). By moving away from traditional values of conserving wickets early on, which is important in the longer forms of the game, teams may be able to increase their chances of winning.

Limitations and future research

It could be argued that the model accuracy of under 70% is still quite low. In this regard, the model could be improved. Future research could include other variables such as the quality of the pitch, and/or the handedness of the batsman and bowler for each ball. Additionally, the type of bowler and type of ball bowled may influence the number of runs scored (e.g., some batsmen may find it easier to score runs from faster bowlers than spin bowlers), and some bowlers may be better suited for bowling at different points of the innings (11). However, for this to occur, more detailed data is needed. Lastly, this analysis did not examine the second inning of matches. It is highly likely that the model accuracy would increase if some measure of bowling performance (following the batting) were also incorporated into the models.

Conclusions

This paper analyzed ball-by-ball data from the first inning of every completed IPL match. It was found that a) teams are at a slight, but significant, disadvantage by batting first, b) total number of runs was unsurprisingly the strongest predictor of match outcome, c) scoring zero or one run from a ball decreases the probability of a win, and d) number of wickets lost is not uniquely predictive of match outcome (although it may have an indirect effect through runs scored). More detailed data is required to increase the accuracy of these machine learning models. That being said, these results are still informative: Teams armed with the knowledge that scoring one run from a ball could theoretically place less emphasis on strike rotation (i.e., scoring one run frequently) and greater emphasis on having batsmen face bowlers in more favorable matchups, thereby likely increasing the number of runs scored in the long-term context of an inning, and being more likely to win the match.

References

1. Sports Show. Top 10 Most Popular Sports in The World. [updated: 2020 October 3; cited 2021 March 7]. Available from https://sportsshow.net/top-10-most-popular-sports-in-the-world/
2. CNBC. Why cricket is worth $5.3 billion – in just one country. [updated: 2018 August 1; cited 2021 March 7]. Available from https://www.cnbc.com/2018/07/03/cricket-ipl-india-sports-mlb-baseball.html
3. Bunker RP, Thabtah F. A machine learning framework for sport result prediction. Applied Computing and Informatics. 2019; 15(1): 27-33. https://doi.org/10.1016/j.aci.2017.09.005
4. Kaggle. IPL Complete Dataset (2008-2020). [updated 2020 November 22; cited 2021 March 7]. Available from: https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020
5. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2021. Available from: https://www.R-project.org/
6. Cricinfo. Records | Twenty20 matches | Team records | Highest innings totals batting second. [updated 2021 March 7; cited 2021 March 7]. Available from: https://stats.espncricinfo.com/ci/content/records/305299.html
7. Wigmore T. Northants' T20 nous a tribute to Moneyball approach [updated 2016 December 16; cited 2021 March 7]. Available from: https://www.espncricinfo.com/story/tim-wigmore-northants-t20-nous-a-tribute-to-moneyball-approach-1070189
8. Kimber J. Where will T20 cricket go next? [updated 2017 May 24; cited 2021 March 7]. Available from: https://www.espncricinfo.com/story/jarrod-kimber-where-will-t20-cricket-go-next-1099221
9. Kapadia K, Abdel-Jaber H, Thabtah F, Hadi W. Sport analytics for cricket game results using machine learning: An experimental study. Applied Computing and Informatics. 2020 Jul 28.
10. Kimber J. Why aren't T20 teams scoring bigger more often? [updated 2019 May 16; cited 2021 March 24]. Available from: https://www.espncricinfo.com/story/jarrod-kimber-why-aren-t-t20-teams-scoring-bigger-more-often-1184438
11. Kimber J. Who do you bowl in the Powerplay in T20? [updated 2018 March 14; cited 2021 March 24]. Available from: https://www.espncricinfo.com/story/jarrod-kimber-who-do-you-bowl-in-the-powerplay-in-t20-1140134