Outcome Prediction in Cricket using pre-game features and in-game scores

George Dockrell
george.dockrell2@mail.dcu.ie

Dublin City University, Collins Ave Ext, Whitehall, Dublin 9, Ireland

Abstract. This project predicts the outcome of a T20 cricket match using a selection of algorithms. The project uses data from the Indian Premier League (IPL) from 2008-2020 to build a pre-game predictive model, and an in-game predictive model. The effects of the toss, home advantage, and the teams involved in each game are investigated. Exploration into certain periods of the game is done to see whether runs scored or wickets lost in the powerplay can identify the match outcome and to what extent. This period is compared with the subsequent overs to understand its relative importance. A selection of algorithms are used and compared with the chosen evaluation metrics. The algorithms used in this project are random forest, logistic regression, naïve Bayes, support vector machine, and XGBoost. The pre-game prediction model resulted in 67% accuracy, the in-game prediction model ranged from 48-65% in the first innings and from 65-89% in the second innings. The in-game model was also evaluated using IPL 2022 data and data from the Big Bash League (BBL) and returned similar results.

Keywords: Binary classification · Machine Learning · Cricket Prediction.

1 Introduction

Outcome prediction is a popular topic that has been explored across a variety of sports. Sport is described as a combination of luck and skill and prediction within sport can prove problematic [1]. Prediction can help in-house personnel with team selection, tactical decisions, and understanding what features contribute to a winning performance. The sports betting industry has driven progress within sport prediction and thrives on the accuracy of predictive models alongside the market's reaction to the odds that have been offered [2].

1.1 T20 Cricket

Cricket has become more data driven thanks to the rise in the T20 format. T20 cricket was introduced in 2003 and is 20 overs long. This gives each side 120 balls to score as many runs as possible which is more easily analysed than multi-day cricket. Additionally, with bigger prize money at stake, many teams are using data to gain a competitive advantage.

1.2 Indian Premier League

The Indian Premier League (IPL) had its first edition in 2008 and has been running for 15 seasons. It is the biggest global T20 league in terms of standard of cricket and revenue. The 2021 season attracted 660 million viewers on television and streaming services and the television and streaming rights are expected to sell for $6-7 billion over the next five years [3].

1.3 Toss

The toss involves the flipping of a coin to determine which captain gets the choice of batting or fielding first. The venue, time of day, opposition strengths and weaknesses, and the strengths and weaknesses of the captains own team may influence batting or fielding first. Morning matches may produce a slightly damp pitch that gets harder and therefore easier to score runs on as the day goes on, while evening games may involve dew which can make it extremely difficult to bowl with a wet ball. Some teams may favour batting or fielding first and so the assumption is that winning the toss gives a team a competitive advantage.

1.4 Home advantage

Another proposed hypothesis is that the home team has a stronger chance of winning than the away team. Some teams have the reputation of having a home ground which is tough for away teams to win at. This is investigated to see if this home advantage is consistent across all teams or only a select few.

1.5 Powerplay

The powerplay (overs 1-6) is seen as a key phase of the game which involves fielding restrictions with two fielders allowed outside a thirty-yard ring to encourage attacking batting. After the powerplay the fielding team can have up to five fielders outside of the ring. Taking wickets within the powerplay can halt a batting team's momentum so it is seen as a crucial period for the fielding side. An in-game predictor's accuracy should improve over time so the hypothesis is that the powerplay should create a rise in accuracy above the normal trend.

This paper seeks to investigate the following hypotheses:

- Toss, and Home advantage improves the pre-game predictive model.
- The powerplay creates a spike in accuracy in the in-game model, compared to the subsequent overs.

2 Related Work

2.1 Prediction within Sport

Statistics have played a part in sporting analysis for many years. Baseball lead from the front thanks to Bill James' annual baseball abstract and sabermetrics

[4]. The baseball community leveraged statistics and analysis, which in many respects showed the potential use of data within sports. Advancements in Machine Learning and improved availability of data have allowed for better insights.
Research into the prediction of National Basketball Association (NBA) results tested a number of classification algorithms with logistic regression yielding the best results followed by neural networks (NNs), support vector machines (SVM) and naïve Bayes [5]. Artificial neural networks (ANNs) have been used in the prediction of football match outcomes in Iranian football [6], horse racing [7], and National Collegiate Athletic Association bowl outcomes [8].

2.2 Prediction within Cricket

One of the earliest uses of statistics within cricket was the Duckworth Lewis method to revise the number of runs required in the case of a rain-reduced match [9]. This sees overs remaining and wickets in hand as resources available to the batting team and can be used to recalculate a relative target that is fair to both teams. This system was later updated to include scoring rates in the calculation of the revised target [10].
Multiple linear regression has been used to predict margin of victory in cricket due to the fact that the difference between scores can be approximated using a normal distribution [11]. In conjunction with the Duckworth Lewis method to convert available resources into runs, multiple linear regression was used with six variables (average margin of victory (MOV) vs opposition, past totals by the batting team, average runs conceded by the bowling team, the home country, the average MOV, and the past totals at the venue) to correctly predict 71% of winning teams in One Day International (ODI) cricket [12]. Other papers have used linear regression to predict the first innings score based on the current score and balls and wickets remaining, and naïve Bayes in the second innings alongside the target score needed with accuracy ranging from 68%-91% depending on the stage of the match [13].
A 74% accuracy was achieved using logistic regression to predict international matches with home team, away team, day/night, International Cricket Council (ICC) ranking, previous form and experienced players stats [14].
Decision trees and multilayer perceptron network were used with the features team 1, team 2, winner, venue, winner type (runs/wickets), and host country with limited success - 55% and 57% respectively [15]. The use of naïve Bayes, decision trees, adaboost and bagging alongside association rule mining and clustering has also been tested with naïve Bayes returning the best results [16].
A novel approach to cricket prediction combined two predictor models to predict boundary and non-boundary balls for the remainder of the game [17]. Using historical features alongside clustering information from batting statistics, boundaries are predicted. Non boundary balls are predicted using ridge regression and combined with boundary balls to predict the final score. Predictions initiated before the start of play use venue information and the batsman cluster scores. This novel method achieved an accuracy of 68%-70% [17], improving upon the results achieved in previous work [12].

Predicting the outcome of ODI cricket using the features - home team, opposition team, toss outcome, bat first innings, and ground location achieved a 60%-61% accuracy using naïve Bayes, SVM and random forest [18].
Previous papers have explored some aspect of pre-game prediction [12, 14–16, 18] and in-game prediction [11, 13, 17, 19, 20]. In-game prediction of ODI cricket has been previously studied, however there are limited studies on the T20 format. Two studies have investigated in-game prediction in the T20 format, but poor methodology resulted in overly accurate results [19, 20]. This project investigates in-game prediction in T20 cricket to predict a win for the team batting first using a more robust methodology. This project also investigates pre-game prediction in order to maximise model accuracy.

3 Methodology

The project was conducted in three parts. 1) Data Analysis was completed to understand the features to be used in the pre-game prediction. 2) Pre-game prediction was completed using the columns that were investigated in the data analysis section and deemed influential through feature selection. 3) In-game prediction was completed using transformed data from the ball-by-ball data set to predict the outcome of the game at various periods using in-game scores. This was further divided into a first innings predictor and a second innings predictor.

3.1 Data

The data set found at IPL Data was used for this project and comprises two separate csv files; matches and ball-by-ball. The matches csv file contains a row for each match played from 2008-2020 and a sample of matches is presented in Figure 1. Ten features were used for the pre-match analysis: city, date, venue, neutral_venue, team1, team2, toss_decision, year, country, and home_won_toss.

	city	date	venue	neutral_venue	team1	team2	toss_decision	year	country	home_won_toss
0	2	0	14	0	11	6	1	2008	0	1
1	7	1	23	0	4	0	0	2008	0	0
2	10	1	8	0	2	9	0	2008	0	0
3	23	2	35	0	7	11	0	2008	0	1
4	22	2	7	0	6	1	0	2008	0	0

Fig. 1. A sample of Matches Data from the IPL data set

The ball-by-ball csv file contains information on each ball bowled across all thirteen seasons of the IPL (2008-2020). The runs scored, and wickets taken off each ball was used for the analysis (Figure 2).

id	inning	over	ball	batsman	non_striker	bowler	batsman_runs	extra_runs	total_runs	non_boundary	is_wicket
335982	1	6	5	RT Ponting	BB McCullum	AA Noffke	1	0	1	0	0
335982	1	6	6	BB McCullum	RT Ponting	AA Noffke	1	0	1	0	0
335982	1	7	1	BB McCullum	RT Ponting	Z Khan	0	0	0	0	0
335982	1	7	2	BB McCullum	RT Ponting	Z Khan	1	0	1	0	0
335982	1	7	3	RT Ponting	BB McCullum	Z Khan	1	0	1	0	0

Fig. 2. A sample of Ball-by-ball Data from the IPL data set

3.2 Algorithms

Five algorithms were used in this paper to observe which returned the best results and to evaluate the strengths and weaknesses of their predictive ability (Table 1).

3.3 Evaluation Metrics

Accuracy, area under curve (AUC), and f1 were the evaluation metrics used in this project for pre-game prediction. Accuracy was used for in-game prediction as it is a simple metric for the proportion of predictions that are correct and has previously been used for the majority of the cricket-related papers [12, 14–18]. AUC was used as it is a more sensitive test compared with overall accuracy [21]. F1 is a balance between precision and recall and therefore is used when there is an uneven class distribution, aiding algorithms with higher sensitivity [22].
Data from the 2010-2021 seasons of the Big Bash League (BBL) in Australia was used to evaluate the adaptability and robustness of the in-game IPL trained model. The results for the IPL 2022 was also used to evaluate the predictive model alongside live odds from a popular betting company (bet365). These odds were noted in-game alongside in-game scores. Although responsive to market bets, odds can often provide a good indication of the likely favourite and so have been used to further evaluate the results of this model. This allows for inspection of inefficiencies or limitations within the betting market.

4 Experiments

4.1 Data Analysis

The IPL matches data set was downloaded and cleaned. Ten relevant features were selected from the data set. Duckworth Lewis affected games, tied games, and no result games were removed from the data set. Each of the ten features

Algorithm	Description
Random Forest	A discriminative model that involves the use of a large number of decision trees and sampling to find the most predicted output.
Logistic Regression	A discriminative model that finds a suitable line separating the two classes and uses the sigmoid function: $$\sigma(z) = 1\frac{1}{1+e^{-}z}$$
Naïve Bayes	This operates under the assumption of independence between all of the features. It is a generative probabilistic model that calculates the conditional probability of each class using Bayes theorem: $$p(A\|B) = \frac{p(B\|A)p(A)}{p(B)}$$
Support Vector Machine	A discriminative model, SVM finds the hyperplane that allows for the splitting of samples into two distinct classes with the largest possible distance between all points and the hyperplane.
XGBoost	XGBoost trains a set of regression trees using gradient descent to minimise loss, and combines these regression trees for its output.

Table 1. Algorithms used

were inspected for any potential insights and the toss and home advantage were investigated at a deeper level. The goal was to understand if there was a historic competitive advantage in winning the toss or playing at home. These two features were inspected with respect to the individual teams and over each season of the IPL.

4.2 Pre-game prediction

The pre-game model predicted a home win or an away win using the cleaned matches data set. This data is categorical and so it was label encoded. A country column, calculated by the city column, was created and added to the data. The 2009, 2014, and 2020 seasons were hosted outside of India and so it was believed that this was important information that may have improved the accuracy of the classifier. Feature selection was used to decide what features were most appropriate for the task, for each algorithm. To improve on this prediction the home and away team columns were one hot encoded and all other columns were removed. Finally, the away team columns were removed, leaving the home team one hot encoded into 13 columns for prediction of a home win. The goal of this

section was to understand how pre-game prediction varied when using all of the data, home and away team data, and with home team data alone.

4.3 In-game Prediction

The in-game prediction looked at various periods within the game to understand how the in-game scores affect the outcome. The ball by ball data was transformed into over by over data to use with the in-game predictive model. The data was divided into first innings data and second innings data. The first innings data contained the over number, team total, and team wickets lost. The second innings data contained the over number, team total, team wickets lost, and the corresponding first innings score and first innings wickets lost at the same point in the game, alongside the first innings final total. This resulted in the second innings data having three extra columns and hence why there were two sets of models. All five algorithms were used for in-game prediction. The hypothesis was that the powerplay score results in a spike in accuracy in comparison to the subsequent overs within the game. The over by over accuracy's were graphically displayed to inspect how prediction changes over time and how prediction changes from 1st innings to 2nd innings.

4.4 Validation

The in-game model wass trained on the IPL data and tested using the BBL data to measure its adaptability across leagues. The IPL 2022 data was used to validate the pre-game model and the in-game model using betting odds. The pre-game odds were collected prior to each game and the in-game odds were collected live during the game. A single row of data was collected during each innings of each game with the score, wickets lost and live odds for both teams. For each row of data a theoretical bet was placed to understand if the model is profitable.

5 Results

5.1 Data Analysis

Toss Captains who won the toss chose to field 60.7% of the time in the IPL. Within professional cricket, when batting second, players can often take more calculated risks and some players are shown to excel in situations where the runs required is known. Upon further investigation there is a clear competitive advantage towards fielding first (Figure 3). When teams chose to field they won 55.4% of their games and when teams chose to bat they won 45.5% of their games. This reinforces the idea that professional cricket teams within the IPL are better at batting second. The trend of teams choosing to field first has also changed over the 13 seasons of IPL with an increase in field first choices especially within the 2016-2019 seasons (Figure 4).

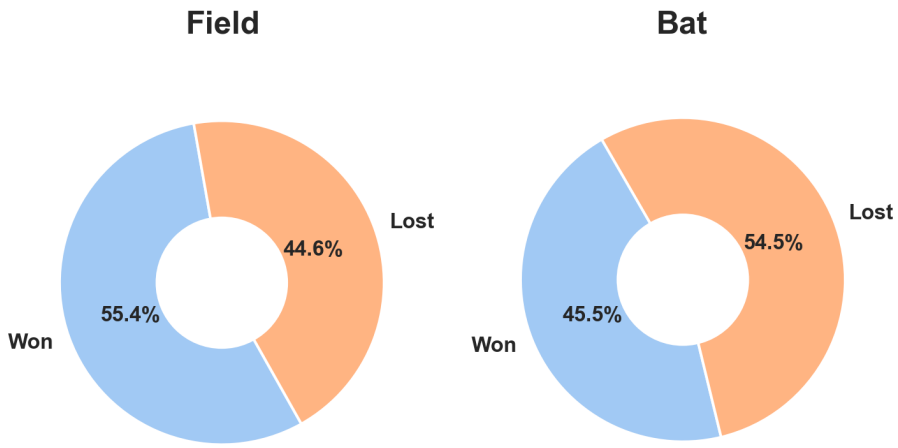

Fig. 3. Winning rate by choice having won the toss

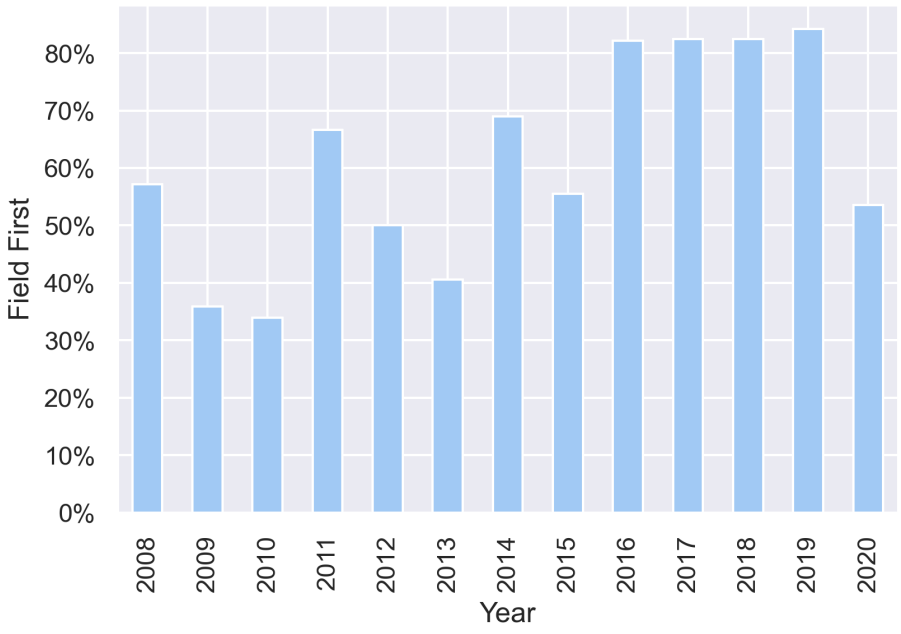

Fig. 4. The decision to field first over time

Home advantage The home team won 55.2% of the games in the IPL. To further investigate this relationship, the win percentage was calculated for all teams at home and away from home. Chennai Super Kings had the strongest win % at home with 70% followed by Mumbai Indians, Sunrisers Hyderabad, and Rajasthan Royals all with 64% home win rate. Mumbai Indians had the strongest away win rate of 59%. Investigating the difference between teams home win rate and away win rate gave an indication of which teams struggled due to travel implications, fan support, or changing conditions (Figure 5). The most notable difference was for the Rajasthan Royals who had an impressive record at home but who struggled when travelling away from home. Chennai Super Kings also performed extremely well at home but struggled when playing away from home. Mumbai Indians showed incredible consistency across all venues and hence have won the most IPL titles. The Gujarat Lions, interestingly performed better when away from home.

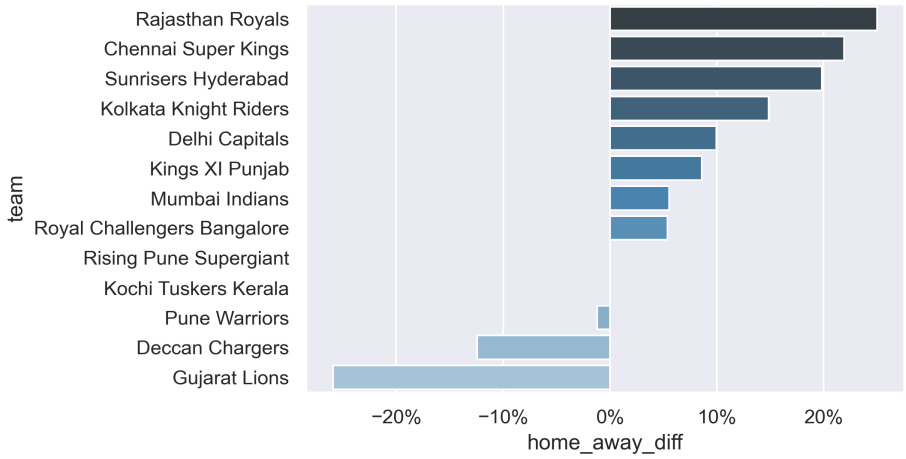

Fig. 5. Difference in Home vs Away win rate

5.2 Pre-Game Prediction

Random forest returned 66% accuracy for pre-game prediction with all 10 features (Table 2). Feature selection was then used to improve upon this (Table 2). It is noted that the feature with highest importance varied by algorithm (random forest - date, logistic regression - toss decision, XGBoost - neutral_venue, SVM - toss_decision). Feature selection resulted in improvements for the logistic regression, naïve Bayes, and XGBoost classifiers but a reduction in the previously highest performing random forest. The final pre-game prediction involved one hot encoding for the home team and away team (Table 2). This resulted

in random forest having the highest accuracy of 65%. The metric was improved upon when the away team column was dropped from the random forest model resulting in a 67% accuracy, the highest of all pre-game predictions. The simplest of all the models returned the best results and showed that the features investigated in the Data Analysis section did not aid prediction any more than the home team feature by itself. All results with multiple evaluation metrics are presented in Appendix A.

Model	**All 10 features**	**Feature Selection**	**One Hot Encoding**
Random Forest	**0.66**	0.57	**0.65**
Logistic Regression	0.47	0.56	0.56
Naïve Bayes	0.58	**0.60**	0.57
Support Vector Machine	0.56	0.55	0.56
XGBoost	0.52	0.57	0.57

Table 2. Pre-Game Prediction Accuracy

5.3 In-Game Prediction

For each innings, the data was grouped by over number and a model was used for each over. This improved the overall accuracy across all models and resulted in a model that was trained/tested on data from the first over of the first innings, across all games, and a separate model for each of the subsequent overs. The in-game accuracy using random forest returned the best results (Figure 6). The accuracy started at less than 50% after the first over of the first innings, and grew to 67% by the end of the first innings. The second innings accuracy started at 65% and grew to 89% at its peak. The largest spikes in accuracy occurred at the 4th over and 18th over in both innings and in the 12th over in the first innings and the 14th over in the second innings. There was a spike after the 4th over but the general trend was upwards all the way through both innings. There was an interesting drop off in accuracy after the 18th over in each innings. This may have been due to the model being unable to give the correct importance to wickets which can quickly turn a game in the final overs. This assumption would need to be further investigated.

5.4 Validation

The BBL data results were similarly successful, showing that the model is applicable to other T20 leagues from around the world.
The evaluation of the system using live in-game betting odds resulted in the successful prediction of 48/62 match outcomes with an accuracy of 77%. The return on investment of these theoretical bets was 28% and although the sample size was small this would indicate the model has some useful predictive capabilities.

The pre-game betting odds correctly predicted 11/24 samples returning a 2.5% loss.

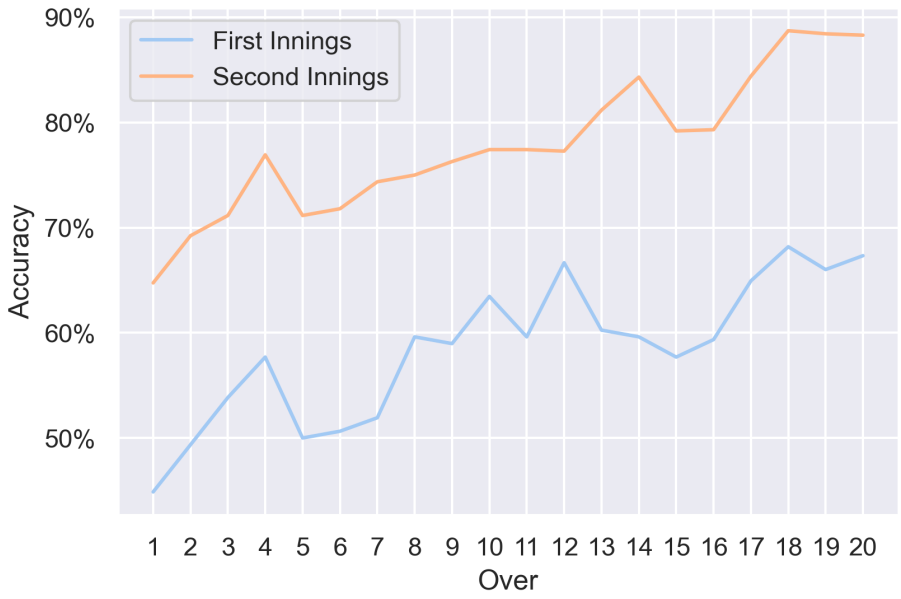

Fig. 6. Over-by-over predictions using Random Forest

6 Findings

The proposed hypothesis that winning the toss gives a competitive advantage to teams was confirmed in the data analysis section. Teams that won the toss and chose to bowl had a competitive advantage over the opposition, however winning the toss and choosing to bat decreased the chances of winning. Home advantage was a factor in winning games but the findings were inconsistent between teams in this respect.

The pre-game prediction using all ten features, a selection of features, home and away data (one hot encoded), and home team data (one hot encoded), all produced similar results. The highest accuracy recorded was for one hot encoding of the home team indicating that the other columns within the dataframe were not as influential on the predicted outcome. This was contrary to the hypothesis. Other work investigating pre-game prediction has returned greater accuracy using player's batting averages and so this may need to be investigated [14, 17].

The in-game prediction model rejected the belief that the powerplay was a greater predictor of success than the subsequent overs. There were dips in accuracy from 18 overs onwards in both innings, contrary to what was believed to

be the easiest periods to predict. The assumption is that the variability around wickets lost at this time in the game and the volatility of T20 cricket in short spaces of time may have hampered the predictions. The second innings prediction was higher than the first innings showing that the ability to reference the corresponding first innings score improved the models predictive accuracy. Both innings prediction accuracy grows over time which is consistent with the belief that as the game progresses the likelihood of one result or another is more apparent.
The random forest classifier was the most effective algorithm across a number of tasks although performance of all algorithms were within a similar range. This utility was across pre-game, using feature selection and one hot encoding, and in-game also. It also performed well with the BBL data and using the in-game IPL 2022 data showing robustness and further applications. The pre-game IPL 2022 validation showed limitations within the model. The 2022 mega auction in which a large number of players changed teams shows that the players and not the teams are responsible for success and that the historic data is of less use with changes in player personnel.

7 Conclusion

This project investigated the IPL data set from 2008-2020. The Data Analysis showed teams are increasingly choosing to field when they win the toss and this is due to the historic competitive advantage given to teams that do so. Contrary to what is often thought, the powerplay overs did not greatly improve the predictive accuracy of the model. The in-game prediction using team totals and wickets lost was successfully applied to the IPL data set as well as the BBL data set. The testing of the model on the 2022 IPL data achieved 77% accuracy and confirmed the predictive capabilities of this model. For future work it would be worthwhile to explore player averages to aid pre-game prediction. It would also be worthwhile to add average scores at a venue to the first innings predictor to see if accuracy could be improved.

References

1. Aoki, R. Y., Assuncao, R. M. & Vaz de Melo, P. O. *Luck is hard to beat: The difficulty of sports prediction* in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada* (2017), 1367–1376.
2. Snowberg, E., Wolfers, J. & Zitzewitz, E. in *Handbook of Economic Forecasting* 657–687 (Elsevier, 2013).
3. Forbes. *Indian Premier League Valuations: Cricket Now Has A Place Among World's Most Valuable Sports Teams* `https://www.forbes.com/sites/mikeozanian/2022/04/26/indian-premier-league-valuations-cricket-now-has-a-place-among-worlds-most-valuable-sports-teams/?sh=1ca18ce43951` (2022).

4. James, B. *The bill james baseball abstract 1987* (Ballantine Books, 1987).
5. Cao, C. *Sports data mining technology used in basketball outcome prediction* MA thesis (Technological University Dublin, 2012).
6. Arabzad, S. M., Tayebi Araghi, M., Sadi-Nezhad, S. & Ghofrani, N. Football match results prediction using artificial neural networks; the case of Iran Pro League. *Journal of Applied Research on Industrial Engineering* **1,** 159–179 (2014).
7. Davoodi, E. & Khanteymoori, A. R. Horse racing prediction using artificial neural networks. *Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing* **2010,** 155–160 (2010).
8. Delen, D., Cogdell, D. & Kasap, N. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting* **28,** 543–552 (2012).
9. Duckworth, F. C. & Lewis, A. J. A fair method for resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society* **49,** 220–227 (1998).
10. Stern, S. E. The Duckworth-Lewis-Stern method: extending the Duckworth-Lewis methodology to deal with modern scoring rates. *Journal of the Operational Research Society* **67,** 1469–1480 (2016).
11. Bailey, M. J. *et al. Predicting sporting outcomes: A statistical approach* PhD thesis (Faculty of Life and Social Sciences, Swinburne University of Technology, 2005).
12. 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,** 480 (2006).
13. Singh, T., Singla, V. & Bhatia, P. *Score and winning prediction in cricket through data mining* in *2015 international conference on soft computing techniques and implementations (ICSCTI), Faridabad, India* (2015), 60–66.
14. Shah, P. & Shah, M. Predicting ODI Cricket Result. *Journal of Tourism, Hospitality and Sports* **5,** 19–20 (2015).
15. Kumar, J., Kumar, R. & Kumar, P. *Outcome prediction of ODI cricket matches using decision trees and MLP networks* in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India* (2018), 343–347.
16. Kaluarachchi, A. & Aparna, S. V. *CricAI: A classification based tool to predict the outcome in ODI cricket* in *2010 Fifth International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka* (2010), 250–255.
17. Sankaranarayanan, V. V., Sattar, J. & Lakshmanan, L. V. *Auto-play: A data mining approach to ODI cricket simulation and prediction* in *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA* (2014), 1064–1072.

18. Pathak, N. & Wadhwa, H. Applications of modern classification techniques to predict the outcome of ODI cricket. *Procedia Computer Science* **87,** 55–60 (2016).
19. Barot, H., Kothari, A., Bide, P., Ahir, B. & Kankaria, R. *Analysis and Prediction for the Indian Premier League* in *2020 International Conference for Emerging Technology (INCET), Belgaum, India* (2020), 1–7.
20. Agrawal, S., Singh, S. P. & Sharma, J. K. *Predicting results of Indian premier league T-20 matches using machine learning* in *2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India* (2018), 67–71.
21. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* **30,** 1145–1159 (1997).
22. Sokolova, M., Japkowicz, N. & Szpakowicz, S. *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation* in *Australasian joint conference on artificial intelligence, Canberra, ACT, Australia* (2006), 1015–1021.

Appendix A

Results

Model	Accuracy	Area Under Curve	Precision	Recall	f1 Score
Random Forest	**0.66**	0.54	0.64	**0.66**	0.58
Logistic Regression	0.47	0.42	0.47	0.47	0.47
Naïve Bayes	0.60	**0.62**	**0.65**	0.60	**0.61**
Support Vector Machine	0.56	0.47	0.51	0.56	0.52
XGBoost	0.52	0.53	0.57	0.52	0.53

Table 3. Pre-Game Prediction with all columns

Model	Accuracy	Area Under Curve	Precision	Recall	f1 Score
Random Forest	0.57	**0.56**	**0.60**	0.57	**0.58**
Logistic Regression	0.56	0.50	0.55	0.56	0.55
Naïve Bayes	**0.60**	0.49	0.53	**0.60**	0.54
Support Vector Machine	0.55	0.50	**0.54**	0.55	0.54
XGBoost	0.52	0.54	0.58	0.52	0.53

Table 4. Pre-Game Prediction using feature selection

Model	Accuracy	Area Under Curve	Precision	Recall	f1 Score
Random Forest	**0.65**	0.57	**0.63**	**0.65**	**0.62**
Logistic Regression	0.56	0.57	0.61	0.56	0.57
Naïve Bayes	0.57	**0.58**	0.62	0.57	0.58
Support Vector Machine	0.56	0.54	0.58	0.56	0.57
XGBoost	0.56	0.56	0.60	0.56	0.57

Table 5. Pre-Game Prediction using One Hot Encoding of home team and away team