CA4021 Project Proposal - Outcome prediction in cricket based on pre-match features and in-game scores

George Dockrell
george.dockrell2@mail.dcu.ie
18745115

Supervisor: Mark Roantree

December 17, 2021

1 Executive Summary

This project proposes to predict the outcome of a T20 cricket match using a selection of features. The project will involve analysis of the Indian Premier League (IPL) from 2008-2020 and will use a readily available dataset from Kaggle to build a predictive model. The effects of the toss, venue, city, and team on the outcome within a match will be investigated. The data will be transformed to create calculated fields for each game. This will allow for further exploration into certain periods of the game to see whether runs scored or wickets lost at various stages can be a useful predictor of the outcome of the game. One focus of the research is whether the score after the six over powerplay can identify the match outcome and to what extent. This period will be compared with the subsequent overs to understand its relative importance. A selection of algorithms will be used and comparisons and contrasts within these will be made using the chosen evaluation metrics. The algorithms are decision tree, random forest, support vector machine (SVM), logistic regression, and naïve Bayes. Please see the Glossary at the end of this document for clarity on cricket specific terms.

2 Motivation and Background

Outcome prediction is a popular topic that has been explored across a variety of sports. Sport is often described as a combination of luck and skill and prediction within sport can often prove problematic (Aoki, Assuncao, and Vaz de Melo, 2017). Prediction within sport can help in-house personnel with team selection, tactical decisions, and understanding what features are most impactful in a winning performance. Another industry that has driven the progress within sport prediction is gambling which thrives on the accuracy of their predictive models alongside the market's reaction to these odds (Snowberg, Wolfers, and Zitzewitz, 2013).

Cricket has become more data driven thanks to the rise in the T20 format. The 120 balls per side is a more easily analysed format than multi-day cricket and with bigger prize money at stake, many teams have turned to Data Analytics to gain a competitive advantage. Cricket is unique as a sport in that the conditions vary greatly from country, city, and venue. The grass wicket that is used for each game can change significantly and so players must adapt to each wicket accordingly. Generally speaking, Australian wickets will be hard and bouncy, English wickets will be green and allow for lateral movement of the ball, and Indian wickets will be slow, low and will spin a lot. This results in varied team selection based on the venue and this variety can also be seen across a selection of venues used in the Indian Premier League (IPL), with each pitch having its own nuances and characteristics. The venue is often an interesting feature as it may be high scoring or low scoring due to the quality of the surface and the distance of the boundaries from the pitch.

Another interesting aspect of a cricket match is the toss. The toss is often seen as a major factor in the outcome of a cricket match. The team that wins the toss can choose to bat or bowl first and this can often dictate whether they get the best conditions during the game. Morning matches may have a slightly damp pitch that gets harder and flatter as the day goes on. Evening games may involve dew which can make it extremely hard to bowl with a wet ball. This is why the toss is seen as playing a major role in the result.

T20 cricket is split into three phases. The powerplay (overs 1-6), the middle overs (7-15) and the death overs (16-20). The first six overs are often seen as playing a major role in the outcome of matches. This period involves certain fielding restrictions that allow for more aggressive batting and can help a batting side to set a platform for a good score. Conversely, taking wickets within the first six overs can greatly halt a batting team's momentum and so it is seen as a crucial period for the fielding side within a T20 match.

3 The Problem Statement

The objective of this project is to predict the outcome of cricket games based on a selection of features. The data used in the predictive model will be from T20 cricket matches and will be exclusively from matches within the IPL from the years 2008-2020. The aim is to gain a greater understanding of what features are most influential in this prediction task. The prediction is a classification problem with the predicted label as a home win or away win. The features used include city, venue, country, toss outcome and teams playing. The full set of features will be further explained in the Methodology section. The project will involve exploring the dataset to identify which aspects of the data carry the strongest weight in prediction and comparing this to domain knowledge within professional cricket as to what is expected. The various periods within the game will be explored to see if there are any trends with respect to what sections of the game have the largest impact on the result. All of this data from pre match features will then be combined with in game scores to see if it can improve the quality of the prediction.
With this three stage approach, it is planned to be able to predict the outcome of games based on the selected features to a reasonable level of accuracy. The goal is to gain some insight into the role the country, city, venue and toss play in a match outcome and also into the various stages of the game and whether certain periods of the game are better predictors of success than others.
This work is based on the IPL dataset and so it is most relevant for T20 cricket played in India. Some seasons of the IPL were played in South Africa, such as in 2009 and 2014 during Indian general elections, and in the United Arab Emirates (UAE) in 2020 due to COVID-19. This allows for some generalisations across our results globally.

4 State of the Art

Statistics have played a part in sporting analysis for many years. Some sports incorporated analytics earlier than others with baseball leading the charge due in large part to Bill James' annual baseball abstract and sabermetrics (James, 1987). The baseball community leveraged statistics and analysis, which in many respects showed the potential use of data within sports. This was most notable when the Oakland A's became a data driven organisation who had great success with a small budget (M. Lewis, 2004). Advancements in Machine Learning (ML) have allowed for better insights as information could improve from summary statistics towards in depth knowledge of predictive features within each individual sport.

4.1 Prediction within Sport

A large variety of sports have used ML in prediction efforts with modern ML algorithms providing sport analysts with larger capabilities of data-driven recommendations. Artificial neural networks (ANNs), decision trees, Bayesian method, logistic regression, support vector machines (SVMs), and fuzzy methods have been used across a variety of sports with varying levels of success (Haghighat et al., 2013).
Research into the prediction of NBA basketball results tested a number of classification algorithms with logistic regression yielding the best results followed by neural networks (NNs), SVM and naïve Bayes (Cao, 2012). ANNs have also been used in the prediction of football match outcomes in Iranian football (Arabzad et al., 2014), horse racing (Davoodi and Khanteymoori, 2010), and NCAA bowl outcomes (Delen, Cogdell, and Kasap, 2012).

4.2 Prediction within Cricket

One of the earliest uses of statistics within cricket was the creation of the Duckworth Lewis method to revise the target in the case of a rain-reduced match (Duckworth and A. J. Lewis, 1998). This system is based on the overs remaining and wickets in hand and sees them as resources available to the team. Therefore, the system can use these as a method to recalculate a relative target that is fair in all situations. This system was later updated to include scoring rates in the calculation of the revised target (Stern, 2016).
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 (M. J. Bailey et al., 2005). In conjunction with the Duckworth Lewis method to convert available resources into runs, multiple linear regression was used with six variables (Ave. margin of victory (MOV) vs opposition, past totals by batting team, ave. runs conceded by bowling team, home country, ave. MOV, past totals at venue) to correctly predict 71% of winning teams in ODI cricket (M. Bailey and Clarke, 2006). Other approaches have used linear regression to predict 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 (T. Singh, Singla, and Bhatia, 2015).
Agrawal, S. P. Singh, and Sharma (2018) predicted match outcomes using SVM, naïve Bayes, and a C tree classifier with the IPL dataset while using created features such as average run rate, average strike rate and powerplay strike rate, hence the accuracy level of 96%, 99% and 98% respectively. Since scoring more runs than the opposition is the method by which games are won, using average run rate creates a feature that will be overly indicative of the winner and therefore would seem to be poor research.
Logistic regression was used to predict international matches with home team, away team, day/night, International Cricket Council (ICC) ranking, previous form and experienced players used with a 74% accuracy (P. Shah and M. Shah,

2015).
Decision trees and multilayer perceptron network were used by J. Kumar, R. Kumar, and P. Kumar (2018) with the features team 1, team 2, winner, venue, winner type(runs/wickets), and host country with limited success - 55% and 57% for decision tree and multilayer perceptron respectively. Kaluarachchi and Aparna (2010) tested the use of classification algorithms such as naïve Bayes, decision trees, adaboost and bagging alongside association rule mining and clustering. It was found that naïve Bayes returned the best results using a set of summary features alongside teams, venue, and toss results.
Sankaranarayanan, Sattar, and Lakshmanan (2014) took a novel approach to cricket analysis using historical data and in match analysis and combining nearest neighbours clustering and linear regression. This novel approach used two different predictor models to predict home runs and non-home run balls for the remainder of the game. It used historical features such as average score, average wickets lost, and average runs conceded and average wickets taken alongside clustering information from five batting statistics. Home run prediction is done using the five nearest neighbours to average the number of home runs achieved in similar situations, historically. Non home runs are then predicted using ridge regression and combined with home runs to predict the final score. Predictions initiated before the start of play use venue information and the batsman cluster scores as a starting point. This novel method improved upon the results achieved by M. Bailey and Clarke (2006) and reduced the level of errors seen. The model developed by Sankaranarayanan, Sattar, and Lakshmanan (2014) achieved an accuracy of 68%-70%.
Pathak and Wadhwa (2016) used naïve Bayes, SVM and random forest to predict the outcome of ODI cricket using the features - home team, opposition team, toss outcome, bat first innings, and ground location and achieved 60%-61% accuracy in the prediction of the results for all three models.
Barot et al. (2020) applied SVM, logistic regression, decision tree, random forest and naive Bayes to the IPL dataset with impressive results ranging from 81.6%(naïve Bayes), decision tree algorithm (87.8%) and logistic regression(95.6%).

5 Methodology

The first step in this problem is to thoroughly inspect the datasets to understand what features will be used and whether any data manipulation will need to occur before it is used in this project's classifiers. Python (Van Rossum and Drake, 2009) and Jupyter notebook (Kluyver et al., 2016) will be used for the data exploration and subsequent prediction models. Scikit-learn (Pedregosa et al., 2011) will be used for these predictive models as it has a selection of algorithms which satisfy our needs for this project. All of the code will be hosted on the DCU School of Computing Gitlab page.

5.1 Dataset

The dataset for this project can be found at IPL Data and comprises two separate csv files, matches and ball-by-ball. The matches csv file contains a row for each match with the following features:

- Match ID
- Neutral Venue
- Winner
- Umpire 1
- Umpire 2
- City
- Team 1
- Team 2
- Result
- Date
- Result Margin
- Method
- Player of the match
- Toss Winner
- Toss Decision
- Eliminator
- Venue

The ball-by-ball csv file contains information on each ball bowled across all thirteen seasons of the IPL. It contains the match ID and runs scored and wickets taken off each ball which we will be using for our analysis.

5.2 Stage one

A country column, calculated by the information that is available on the cities for each match, will be created. The 2009, 2014, and 2020 seasons were hosted outside of India and so it is believed that this is important information that will improve the accuracy of the classifier. Once the country column has been calculated these features will be used to predict the winner for each match. The selection of algorithms used to do so are decision tree, random forest, support vector machine (SVM), logistic regression, and naïve Bayes. The weightings applied to the various columns will be inspected to see what features are the top predictors of the winner. The results of this classification will be cross referenced with domain knowledge of features to see if there are any interesting insights or anomalies.

5.3 Stage two

The second stage of the project involves the second file (ball-by-ball) which has information on every ball from every game between 2008-2020 in the IPL. The aim is to extract features from this file to add to the matches data. Within T20 cricket there are twenty overs comprising six balls each. The first six overs are called the powerplay with only two fielders allowed outside a thirty yard ring to allow for more attacking batting (Figure 1). After this six-over period the fielding team is allowed to have up to five members of the team outside of the thirty yard ring and so the first six overs is often a large factor in a teams success. The second stage of this project aims to look at various stages within the game and to understand how the in-game scores affect the outcome. A team

score and wickets lost will be calculated. Columns will be created for the team total and wickets lost at each over from overs 6-20. The hypothesis is that the score should become a better predictor of success as time goes on so we are interested to see whether the first six overs are a better predictor than overs 7, 8, 9 etc. Wickets lost are often another feature which affect the scoring rate of a team. The wickets lost after six overs is an interesting feature which is hoped to be a strong predictor of success. This will be calculated as a feature to see if it improves the accuracy of our predictive model.

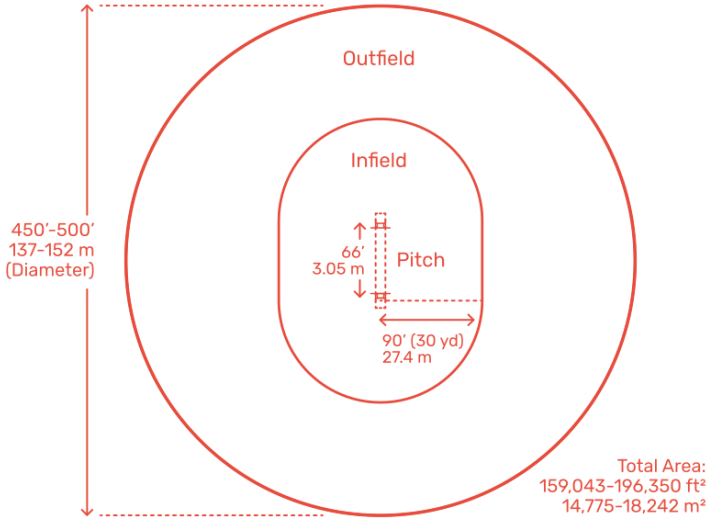

Figure 1: Cricket Pitch Dimensions (*Cricket Ground Dimensions* 2021)

5.4 Stage three

The third stage of this project involves combining the data from the first two stages to understand if this improves the accuracy of our prediction classifier. The methods described in the State of the Art section have used a selection of pre-match features for prediction (P. Shah and M. Shah, 2015; J. Kumar, R. Kumar, and P. Kumar, 2018; Kaluarachchi and Aparna, 2010; Pathak and Wadhwa, 2016), or in-game scores for prediction (T. Singh, Singla, and Bhatia, 2015). In-game scores combined with summary features for each match has been explored but only within 50 over cricket (M. Bailey and Clarke, 2006; Sankaranarayanan, Sattar, and Lakshmanan, 2014) which can be very different to the T20 format. It is hoped the technique of using match features alongside in-game statistics within T20 cricket will improve upon the accuracy of these previous results.

5.5 Evaluation

Accuracy, area under curve (AUC), and f1 are the evaluation metrics. Accuracy is chosen as a simple metric for how many predictions are correct and has previously been used for the majority of the cricket related papers (M. Bailey and Clarke, 2006; P. Shah and M. Shah, 2015; J. Kumar, R. Kumar, and P. Kumar, 2018; Kaluarachchi and Aparna, 2010; Sankaranarayanan, Sattar, and Lakshmanan, 2014; Pathak and Wadhwa, 2016; Barot et al., 2020). AUC can be a good metric to use as it is a more sensitive test compared with overall accuracy (Bradley, 1997). F1 is a balance between precision and recall and therefore can be of use when there is an uneven class distribution, aiding algorithms with higher sensitivity (Sokolova, Japkowicz, and Szpakowicz, 2006). The dataset will be split 80/20 between training and testing to aid evaluation.

6 Project Plan

The project plan contains seven sections. Data Analysis will commence 19th December 2021, followed by the first, second, and third round of experiments. Round 1 will involve prediction using summary features found in the matches file. Round 2 involves prediction using a data transformation from the ball-by-ball file into calculated features. The third and final round will combine these two sets of features to run predictive analysis. Each section is reliant on the completion of the prior section before it can commence. The project paper will be continually edited as each section is completed. The paper will be submitted the first week in May 2022 which will result in seventeen weeks of work from start to finish. This information is detailed in the Gantt chart (Figure 2). The DCU semester two weeks can also be seen above the months in Figure 2. Weeks 8-10 tend to be a very busy time during the semester and this coincides with the final stage of experiments. Four weeks have been allocated to finish the paper, once the final experiments have been completed, and in the case that any of this work is delayed this final write up period may be shortened.

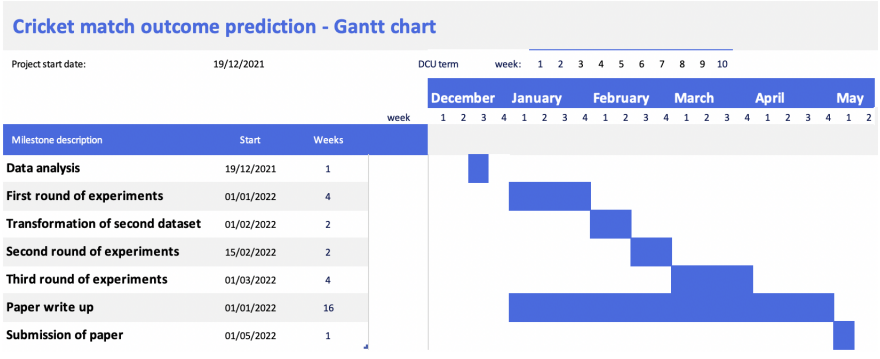

Figure 2: Project timelines

References

Agrawal, Shilpi, Suraj Pal Singh, and Jayash Kumar Sharma (2018). "Predicting results of Indian premier league T-20 matches using machine learning". In: *2018 8th International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, pp. 67–71.

Aoki, Raquel YS, Renato M Assuncao, and Pedro OS Vaz de Melo (2017). "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*, pp. 1367–1376.

Arabzad, S Mohammad et al. (2014). "Football match results prediction using artificial neural networks; the case of Iran Pro League". In: *Journal of Applied Research on Industrial Engineering* 1.3, pp. 159–179.

Bailey, Michael and Stephen R Clarke (2006). "Predicting the match outcome in one day international cricket matches, while the game is in progress". In: *Journal of sports science & medicine* 5.4, p. 480.

Bailey, Michael J et al. (2005). "Predicting sporting outcomes: A statistical approach". PhD thesis. Faculty of Life and Social Sciences, Swinburne University of Technology.

Barot, Harshit et al. (2020). "Analysis and Prediction for the Indian Premier League". In: *2020 International Conference for Emerging Technology (INCET)*. IEEE, pp. 1–7.

Bradley, Andrew P (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7, pp. 1145–1159.

Cao, Chenjie (2012). "Sports data mining technology used in basketball outcome prediction". In.

Cricket Ground Dimensions (2021). URL: `https://www.dimensions.com/element/cricket-ground` (visited on 12/15/2021).

Davoodi, Elnaz and Ali Reza Khanteymoori (2010). "Horse racing prediction using artificial neural networks". In: *Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing* 2010, pp. 155–160.

Delen, Dursun, Douglas Cogdell, and Nihat Kasap (2012). "A comparative analysis of data mining methods in predicting NCAA bowl outcomes". In: *International Journal of Forecasting* 28.2, pp. 543–552.

Duckworth, Frank C and Anthony J Lewis (1998). "A fair method for resetting the target in interrupted one-day cricket matches". In: *Journal of the Operational Research Society* 49.3, pp. 220–227.

Haghighat, Maral et al. (2013). "A review of data mining techniques for result prediction in sports". In: *Advances in Computer Science: an International Journal* 2.5, pp. 7–12.

James, Bill (1987). *The bill james baseball abstract 1987*. Ballantine Books.

Kaluarachchi, Amal and S Varde Aparna (2010). "CricAI: A classification based tool to predict the outcome in ODI cricket". In: *2010 Fifth International Conference on Information and Automation for Sustainability*. IEEE, pp. 250–255.

Kluyver, Thomas et al. (2016). "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press, pp. 87–90.

Kumar, Jalaz, Rajeev Kumar, and Pushpender Kumar (2018). "Outcome prediction of ODI cricket matches using decision trees and MLP networks". In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. IEEE, pp. 343–347.

Lewis, Michael (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Pathak, Neeraj and Hardik Wadhwa (2016). "Applications of modern classification techniques to predict the outcome of ODI cricket". In: *Procedia Computer Science* 87, pp. 55–60.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Sankaranarayanan, Vignesh Veppur, Junaed Sattar, and Laks VS Lakshmanan (2014). "Auto-play: A data mining approach to ODI cricket simulation and prediction". In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, pp. 1064–1072.

Shah, Parag and Mitesh Shah (2015). "Predicting ODI Cricket Result". In: *Journal of Tourism, Hospitality and Sports* 5, pp. 19–20.

Singh, Tejinder, Vishal Singla, and Parteek Bhatia (2015). "Score and winning prediction in cricket through data mining". In: *2015 international conference on soft computing techniques and implementations (ICSCTI)*. IEEE, pp. 60–66.

Snowberg, Erik, Justin Wolfers, and Eric Zitzewitz (2013). "Prediction markets for economic forecasting". In: *Handbook of Economic Forecasting*. Vol. 2. Elsevier, pp. 657–687.

Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz (2006). "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation". In: *Australasian joint conference on artificial intelligence*. Springer, pp. 1015–1021.

Stern, Steven E (2016). "The Duckworth-Lewis-Stern method: extending the Duckworth-Lewis methodology to deal with modern scoring rates". In: *Journal of the Operational Research Society* 67.12, pp. 1469–1480.

Van Rossum, Guido and Fred L. Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. ISBN: 1441412697.

Glossary

Aggressive batting. A batter looking to maximize scoring by swinging harder and looking for a four or six each ball.
Boundary. The marker for the edge of play. Often marked with a rope it can vary from venue to venue from 60-100 meters from the center of the pitch. Boundary can also refer to a Four or a Six.
Death. This refers to the last 30 balls of the innings where batters try to get as many runs as possible and often results in aggressive batting and defensive bowling
Format. This is the length of the game. It can vary from T10(60 balls), T20(120 balls, 50 over(300 balls) or a selection of longer format games (2/3/4/5 days).
Four. If the ball is hit by the batter and crosses the boundary having hit the ground at all it is a four and four runs are awarded to the batting side
ICC ranking. A ranking system calculated by an algorithm that rewards victories vs higher ranked opposition
Innings. The batting turn from one team. It can also refer to the batting turn of a single player. T20 cricket is split into two innings (one for each team)
Over. Six balls in a row from one bowler from one end of the pitch
Power-play. The first 30 balls (overs 1-6) where the fielding team are allowed two fielders outside of the thirty yard ring (Figure 1)
Pitch. The grass that each game is played on. This can vary greatly from country to country and is usually prepared by cutting the grass very short and rolling the wicket with a heavy roller to compact the surface
Runs. A team is rewarded with a run if both batters run the length of the pitch and make their ground. They can run for one, two, three, or four+ runs(very rare) or they can hit the ball for a four or six to get these runs.
Six. If the ball is hit by the batter and travels over the boundary without hitting the ground it is a six and six runs are awarded to the batting side.
Toss. A coin toss to decide which team chooses to bat or bowl first
T20. A cricket match where both sides face a maximum of 120 balls each
Wicket. This can refer to the grass which the game is played on. It is also used to describe when a player is out.