Expected Economy Rate

Nirodha Epasinghege Dona, Robert Nguyen, Paramjit Gill and Tim Swartz *

Abstract

This paper introduces the expected goals concept to limited overs cricket where ideas are illustrated using the economy rate statistic. The approach is primarily explored as a proof of concept since the detailed data that are required for full adoption of the proposed methods are not currently widely available. The approach is based on the estimation of batting outcome probabilities given detailed data on each ball that is bowled in a match. Machine learning techniques are used for the estimation procedure.

Keywords : ball-by-ball data, expected goals, feature identification, model validation, random forests, supervised learning.

*N. Epasinghege Dona is a PhD candidate and T. Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6, R. Nguyen is a PhD candidate, Department of Statistics, School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052, Australia, and PĠill is Associate Professor, Mathematics, Statistics and Physics Unit, Irving K. Barber School of Arts and Sciences, University of British Columbia Okanagan, 3333 University Way, Kelowna BC, Canada V1V1V7. Swartz has been partially supported by the Natural Sciences and Engineering Research Council of Canada. The authors thank Cricket Australia for access to their datasets and the Canadian Statistical Sciences Institute (CANSSI) Collaborative Research Team (CRT) in Sports Analytics who also partially supported the research. The authors also thank an anonymous reviewer whose comments helped improve the paper.

1 INTRODUCTION

Expected goals (xG) is a concept that has gained rapid adoption in professional sport, particularly in soccer. The statistic xG attempts to quantify what is most likely to have happened in a match given the opportunities that occurred during the match. For example, imagine that Team A drew with Team B with the scoreline 1-1 but the expected goals for the match was 3.8 to 0.5 in favour of Team A. In this case, one would conclude that Team A outplayed Team B, and that Team B was fortunate to achieve the draw. Therefore, xG provides a measure of dominance in matches, and attempts to remove what might be described as the "luck" element of sport. Not only does xG describe match dominance, but xG is predictive of future results.

Although xG is intuitive, its calculation is viewed as a black-box procedure from the point of view of the general public. For this reason, and for the technical underpinnings used in the calculation of expected goals, xG falls under the category of advanced analytics. Further, the "black-boxness" of xG is magnified since there are many implementations of xG, and these implementations are typically proprietary. In soccer and hockey, the basic idea behind xG is that there are scoring opportunities on the field and the rink, respectively. The scoring opportunities have associated goal scoring probabilities where goals resulting from shots that are further away and taken from more extreme angles are less probable. These probabilites are summed over the opportunities for each team, leading to the team's xG. The literature that does exist concerning xG is mostly found on blog sites, Twitter feeds and conference proceedings. Some of the more detailed contributions related to xG include Pollard, Ensum and Taylor (2004), Rudd (2011), Macdonald (2012), Decroos et al. (2018) and Fernández, Bornn, and Cervonne (2019).

In this paper, we introduce the the concept of xG to limited overs cricket in the context of the economy rate statistic. In limited overs cricket, the economy rate for a bowler is defined as the average number of runs conceded per over where there are six balls per over. The average is typically calculated over a match, a series, a year or a career. For illustration, we consider economy rate based on a match in which case the economy rate for a bowler is defined as 6*(the number of runs conceded in a match) divided by the number of balls bowled by the bowler in the match. Smaller values of the economy rate are indicative of good bowling.

The development of xER (expected economy rate) is conceptually simple. Based on the characteristics of a ball that has been bowled, we estimate the probabilities of the 8 batting outcomes: wicket, 0 runs, 1 run, 2 runs, 3 runs, 4 runs, 5 runs and 6 runs where we note that there is negligible probability of scoring 3 runs and 5 runs. Using obvious notation, we define the *expected economy rate* in a match for a bowler as

$$\text{xER} = (6/M) \sum_{i=1}^{M} \left(e_i + p_i(1) + 2p_i(2) + 3p_i(3) + 4p_i(4) + 5p_i(5) + 6p_i(6)\right) \tag{1}$$

where i corresponds to the ball number, e_i is the actual number of extras accumulated on the ith ball and M is the number of balls bowled by the bowler in the match. Analogous to xG, xER represents the expected economy rate performance of the bowler in the match. Whereas the p_i's in (1) have been estimated, we do not estimate extras. We consider the observed extras e_i as penalty terms that are directly attributable to the bowler and are added to the expected economy rate formula. There is no luck aspect associated with extras; bad balls are simply bad balls.

Now, xER will only be informative and useful provided that the estimated probabilities p_i in (1) are realistic. To motivate the approach, suppose that $p_i(4) = 0.5$ and the actual result of the bowled ball was a wicket. What might have happened in this scenario is that the bowler's delivery ought to have been exploited by the batsmen. Based on the characteristics of how the ball was bowled, the probability of scoring four runs was high. With a ball of this type, the bowler is typically punished. Instead, the bowler was "lucky" in the sense that a wicket occurred. For example, perhaps a fabulous catch was made. Therefore xER is an attempt to represent what the bowler would have achieved under ordinary circumstances given the performance. The actual achievement is subject to both performance and the luck/stochastic element of sport.

The utility of the xER statistic (1) occurs when a bowler's actual economy rate differs considerably from their xER statistic. For example, suppose that there is a relatively young bowler whose actual economy rate is 8.4 but their xER = 6.6. This would signal that the bowler has been unlucky, and that this is a player for whom team selectors should give attention. This may be a promising bowler.

The elimination of "luck" from performance is a driving force in this research. Whereas luck does not seem to have been addressed in cricket, luck has been investigated in other sports analytics research. For example, luck has been explored in soccer (Sarkar and Kamath 2022), golf (Connolly and Rendleman 2008), baseball (Bailey, Loeppky and Swartz 2020), and hockey (Weissbock 2014).

In this project the probability estimates p_i in (1) are based on detailed ball-by-ball data. Ball-by-ball data have been utilized in various research initiatives in cricket (Swartz 2017). In these projects, ball-by-ball results and covariates have been parsed from match commentaries provided by www.cricinfo.com. However, the ball-by-ball data used in these investigations are not sufficiently detailed for the current investigation. In this project, Cricket Australia has provided us with even more detailed ball-by-ball data from the Twenty20 (T20) format. However, such data are not widely available and do not currently exist for competitions outside of the Australian context. For this reason, the work presented here is explored as a proof of concept. We demonstrate that the approach is feasible and informative, and is something that can be fully developed when detailed ball-by-ball data become widespread. With the advent of player tracking data and analyses across major sports (Gudmundsson and Horton 2017), we expect that the availability of detailed data in cricket is only a matter of time.

Clearly, xER can be extended to applications over series, matches, years and careers. Also, it is clear that other standard bowling statistics such as bowling average and bowling strike rate have xG adaptations. With respect to batting, we might similarly define xG statistics corresponding to batting average and the batting strike rate. Many statistics have been developed for the sport of cricket; see Swartz (2017) for a review of various measures of player evaluation.

In Section 2, we describe the detailed data that have been provided by Cricket Australia. Subjective decisions on retaining and excluding variables are part of the feature identification process. Whereas the focus of sports analytics research has typically involved "big" sports that involve male participation, a feature of this work is that we also analyze women's T20 data. This has the added benefit of assessing differences in how T20 is played between the men's and women's game. In Section 3, we provide a description of the random forest approach used to estimate the probabilities in (1). The procedure falls under the topic of supervised learning. In Section 4, we investigate the quality of estimation in various ways. For example, we examine the overall rate of correct predictions, we examine the related confusion matrices and we qualitatively assess the variables of importance. We also compare our random forest predictions against predictions made by two other methods using Brier scores. We observe that the proposed estimation technique is superior to the other methods. In Section 5, we apply the approach to datasets involving T20 cricket matches. Some insights are obtained which demonstrate the potential of utilizing xER in player evaluation. Finally, a short discussion is provided in Section 6.

2 DATA

Detailed ball-by-ball data have been collected by Cricket Australia over the years 2007 through 2019. The data correspond to international matches involving the Australian national teams (men and women) in Test, ODI and T20 formats. Data have also been collected for Australian domestic matches such as the Big Bash competition. For illustration, we have restricted our attention to T20 matches. There are 532 matches for the men and 725 matches for the women. For a given match, there are an astounding 360 variables collected on every ball that is bowled. The data were coded manually by data entry specialists who watch video broadcasts of matches. It is believed that the data have a high level of accuracy. After some minor data management, our dataset of T20 matches consist of 123,067 bowled balls (men) and 166,898 bowled balls (women) that occurred during the first and second innings.

In our investigation of xER, we are interested in bowling performance. Hence we seek variables that relate bowling performance to the batting outcome. The beauty of sport is that domain knowledge is often high, and feature selection may be assisted by this subjective knowledge. In Table 1, we list $k = 25$ covariates (features) that we believe are predictive of the batting outcome. We have taken the point of view to err on the generous side and include all variables that may have a chance of improving supervised learning; machine learning algorithms have been designed to detect the most important variables. Note that the variables are categorized as either bowling variables or batting variables. A bowling variable is one that is entirely due to the bowler or the conditions of the match. For example, the *ball speed* variable is a bowling variable. As another example, the *wickets lost* variable is a bowling variable. A batting variable concerns something that the batsman did. For example, *batsman handedness* is a bowling variable (because this is a condition of the match) whereas *hit angle* of the batted ball is a batting variable.

In Table 1, we have added two variables that were not included in the Cricket Australia database which we believe are predictive of the batting outcome. The first is the *resources remaining* variable (Duckworth and Lewis 1998) adapted for T20. The resources remaining at the time that a ball is bowled provides an indication of how aggressive a batsman may bat. To also account for batting aggression, we include the *deficit* variable which is the number of runs by which the batting team is trailing in the second innings (i.e. deficit = target score less current batting score). In the first innings, we define *deficit* as the average runs scored in the first innings of a T20 match less the current score. In men's T20 cricket, the average first innings score is 160 runs whereas in women's T20 cricket, the average first innings score is 131 runs.

In Table 1, we observe some redundancies in the variables. For example, the D/L *resources remaining* variable is a function of *wickets lost* and *overs*. In theory, the specification of redundant variables is unnecessary since machine learning techniques are "smart" and are able to detect functional relationships. However, in our experience, over-specification of variables is sometimes useful to assist the performance of algorithms. We also note that the *hit angle* variable has been standardized to account for lefthanded and righthanded batsmen. There were some very minor data management issues such as correcting entries with 10, 11 or 12 wickets lost. These are obvious coding errors where the correct values can be imputed by looking at the data corresponding to adjacent balls.

3 RANDOM FORESTS

Recall that our problem involves the estimation of the batting outcome probabilities $p_i(1), \ldots, p_i(6)$ in (1). These probabilities are estimated in a supervised learning context the batting outcomes and the features (Table 1) are known for each ball in our massive dataset.

A rationale for machine learning methods in prediction is that complex phenomenon are often difficult to model explicitly. Here, we have a categorical response variable y with 8 categories, and a moderate-dimensional explanatory vector $x = (x_1, x_2, \ldots, x_k)$, with $k = 25$. We have little apriori knowledge about the relationship between y and x. For example, the relationship may only involve a subset of the variables x, the components of x may be correlated, and most importantly, the relationship $y \approx f(x)$ involves an unknown and possibly complex function f. In addition, the stochastic aspect of the relationship is typically unknown and big data sets may introduce computational challenges. Miraculously, machine algorithms provide black box predictions based on the features of interest.

For this application, we use random forests as the chosen machine learning algorithm. Random forests (Genuer and Poggi 2020) are particularly easy to implement using the *randomForest* package (Liaw and Wiener 2002) in the R programming language. The basic idea is that a random forest is a collection of many decision trees where prediction results are aggregrated over trees. The use of multiple trees improves prediction and makes inference less reliant on a single tree. The splits in the trees accommodate non-linear relationships and terminal nodes provide the estimated probabilities $p_i(1), \ldots, p_i(6)$.

In choosing the tuning parameters, we have a preference for simpler models (i.e. smaller trees). For example, if a more expansive tree has similar prediction accuracy to a modest tree, we choose the modest tree. To assess accuracy (Section 4), the data were randomly divided where 20% of the observations were used for training and the remaining 80% of the observations were used for validation and prediction. The 20/80 ratio is a little low compared to many applications. However, we want a large validation (prediction) set so that there are enough balls to reliably estimate xER for many of the bowlers.

With 20% of the data restricted to training, this still provides a large enough dataset to obtain a good model. When modifying the training set from 20% to 50% of the observations, we found little change in predictions. In the training component, 10-fold cross-validation was utilized, and this is what allowed us to set tuning parameters. For example, the number of variables randomly selected at each split was set at mtry = 10 to maximize accuracy. We specified 500 trees in the random forest which is the default value. We also used default values for tree depth and the maximum number of nodes. For missing values in our dataset (of which there are few - see Table 1), we chose the argument *na.roughfix* which involves a simple imputation scheme.

4 MODEL VALIDATION

This section investigates the quality of estimation in various ways.

Their are four analyses that are of interest. First, we have the T20 data divided into the men's game and the women's game. And then, within each of these two formats, we consider an analysis A based only on both bowling and batting features (see Table 1). And then we consider an analysis B based only on bowling features (see Table 1). The appeal of Analysis A is that it includes the variables *hit angle* and *hit length*. One might presume that these are very predictive features, which provide more accurate probabilities $p_i(1), \ldots, p_i(6)$, and hence, better estimates of xER. The appeal of analysis B is that it removes the quality of the opposition from the analysis. For example, suppose you have a bowler who only competes against inferior opponents. Then this bowler's observed economy rate would be lower than if the bowler competed against more challenging competition. But this would not be a problem for xER (under Analysis B) since only the characteristics of the bowled ball and the state of the match are considered; the quality of the opposition is eliminated from the evaluation of xER.

4.1 Confusion Matrices

Once a model is trained (fitted), we consider each ball from the validation set. The features of the ball are fed to the model and the batting probabilities are estimated. The batting outcome with the maximum probability is considered the predicted outcome.

For each ball in the validation set, we compare the actual batting outcome with the predicted outcome, and summarize the results in a confusion matrix. The entry (i, j) in the confusion matrix records the number of times an actual batting outcome j was predicted as outcome i.

Tables 2, 3, 4 and 5 provide the confusion matrices for Analysis A (Men), Analysis B (Men), Analysis A (Women) and Analysis B (Women), respectively. The first thing that can be calculated from the confusion matrices is that the overall percentage rates of accuracy for the four analyses are 76%, 45%, 78%, and 47%, respectively. Therefore, as anticipated, the batting features *angle* and *hit_length* greatly assist in the accuracy of the predictions. In Analysis B, one prediction that is particularly poor concerns wickets. Although the wicket calculation does not directly appear in the xER formula (1), its underestimation impacts the other batting outcomes. For example, in Table 3, only 124 wickets were predicted in the roughly 5000 cases where wickets actually occurred. The prediction of wickets is much improved in Table 2 (Analysis A). The same comment also applies to the women's game. In all analyses, we observe that 3's and 5's are almost never predicted. In fact, this is sensible, as they almost always result from some sort of fielding error.

One of the important observations distinguishing Tables 1 and 2 (men) from Tables 3 and 4 (women) is the rate at which 6's occur and are predicted. In the men's game, 6's occur $4123/98450 \rightarrow 4.2\%$ of the time whereas in the women's game, 6's occur $1654/133515 \rightarrow 1.2\%$ of the time. This is a reminder that there are significant differences between the two formats of cricket and that they should be studied separately.

4.2 Features of Importance

One of the informative outputs from the *randomForest* package (Liaw and Wiener 2002) are importance plots of the model features. Features are listed from top to bottom according to their impact on prediction.

We provide importance plots for the men's game (Figure 1) and for the women's game (Figure 2). For each format, importance plots are provided for Analysis A (bowling and batting features) and Analysis B (bowling features only). As expected, we observe that the variables *hit length* and *hit angle* are the most influential as they describe the characteristics of the batted ball. For example, balls that are hit far are most likely to result in runs scored.

In both Analysis A and Analysis B, we observe that the landing location of the bowling delivery (*pitchx* and *pitchy*) impact the batting outcome prediction. The same is true for (*batsmanx* and *batsmany*) which describe the landing location of the ball relative to the batsman. For example, if the ball is too close to the batsman, the batsman is unable to apply full torque on the batted ball, and is less likely to generate 4's and 6's. Our introduced variables *deficit* and *resources remaining* which together describe the urgency and aggressiveness of batting are also variables which help the prediction of batting outcomes.

Comparing Figure 1 and Figure 2, we do not observe many meaningful differences between the men's game and the women's game. It appears that the features that are influential in predicting batting outcomes are similar between the two formats.

4.3 Brier Scores

As mentioned previously, the utility of xER is only as good as the reliability of the estimates of the probabilities p_i in (1). To assess the estimates, we calculate Brier scores (Brier 1950) based on three forecasts.

The first forecast is naive and is not expected to be accurate. However, it does provide a sense of the magnitude of differences with respect to Brier scores. We refer to the first forecast as Uniform Discrete where the associated probabilities are given in Table 6. Here, we set all of the six non-neglible probabilites in (1) equal to 1/6. We emphasize that the probabilities are the same for every ball that is bowled.

The second forecast which we refer to as T20 Proportions is based on the observed proportions of T20 batting outcomes from a much larger dataset. It includes more than 500,000 balls from international men's T20 matches from 2015-2020. Again, we assign zero probability to the rare batting events corresponding to 3 and 5 runs. The probabilities are given in Table 6. Since these probabilities are associated with the men's game, our Brier score analysis will only consider batting predictions for men.

Now, for every ball $i = 1, \ldots, N$ that is bowled in the dataset, we define $o_i(j) = 1$ if event j occurred and $o_i(j) = 0$ if event j did not occur where the event j takes on the values $w, 0, \ldots, 6$. With event probabilites $p_i(j)$ corresponding to a particular forecasting method, the Brier score is then given by

$$B = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=w,0,\ldots,6} \left(p_i(j) - o_i(j)\right)^2 \ . \tag{2}$$

For the men's game, we calculate Brier scores using (2) for the Uniform Discrete forecast, the T20 Proportions forecast and the random forests forecast (Analysis A), We obtain Brier scores of 0.84, 0.73 and 0.33, respectively. The results suggest that the random forests algorithm (Analysis A) predicts the batting outcome much better than the other two methods that do not consider the circumstances of the match and the characteristics of the ball that was bowled. In turn, and most importantly, this suggests that xER informs us about what might reasonably have happened in matches involving economy rate had the batting outcomes proceeded as expected.

5 RESULTS - xER in T20 Cricket

For a given bowler, we are interested in the comparison of xER with their career economy rate. Career economy rates in T20 were obtained from the stats.espncricinfo.com website.

For this analysis, we are only interested in bowlers who have bowled sufficiently in our validation (prediction) dataset. We therefore restricted our attention to bowlers who have bowled at least 500 balls in the validation set. From these bowlers (57 men and 89 women), we randomly selected 10 men and 10 women and calculated their corresponding xER. The calculation was based on the estimates $p_i(1), \ldots, p_i(6)$ in (1) for each ball i that was bowled in the validation set. The xER statistic in (1) was calculated two ways; using Analysis A which relied on bowling and batting features, and Analysis B which only used bowling features. The results are presented in Table 7 (men) and Table 8 (women).

Our initial observation is that the expected economy rates xER are in line with the career economy rates. Sometimes the xER statistic is smaller (the bowler has been slightly unlucky in actual matches) and sometimes the xER statistic is larger (the bowler has been slightly lucky in actual matches). We have argued up to this point that Analysis A is most likely better than Analysis B since Analysis A has more accurate predictions. Perhaps the most interesting bowler according to Analysis A is Renee Chappel from Table 8. Her xER (Analysis A) is a full 2.81 runs lower than her career economy rate. This is a large difference which suggests that she is a better bowler than her record indicates. She is an experienced bowler, 38 years of age and having made her international debut in 2013. In 2016-17, Chappel received the Karen Read Medal as the best player in WACA Female A Grade and T20 competitions. It appears that her excellence as highlighted by the xER statistic has been appreciated.

We also observe that the men bowlers tend to have higher economy rates on average than the women bowlers. This could be related to the observation from Section 4.1 where we observed that the men score 6's more frequently than the women.

6 DISCUSSION

This paper introduces expected economy rate xER to the sport of cricket. The idea borrows on the expected goals concept which has become especially popular in soccer. As xER attempts to reduce the luck element from bowling, xER may be a diagnostic that informs us of the true quality of a bowler, perhaps a more trustworthy statistic than the actual economy rate. This may be particularly valuable in the context of unproven bowlers who have not established a clear reputation. For example, xER could alert team selectors to promising bowlers whose results have not yet matched their quality.

This paper is intended to be a proof of concept of a potentially valuable statistic. Naturally, the statistic could improve with better data. For example, the positioning of fielders surely has an impact on batting outcomes. We suggest that the availability of such data is not far down the road as player tracking data becomes more widely available.

Of course, different models and prediction schemes could also be investigated. This is a possible avenue for future research. Another point of reference for future research in cricket analytics is a call for more work on the women's game. This research has pointed out that there are differences between the men's and women's games.

7 REFERENCES

Bailey, S.R., Loeppky, J. and Swartz, T.B. (2020). The prediction of batting averages in Major League Baseball. *Stats*, 9, 84-93.

Brier, G.W. (1950). "Verification of forecasts expressed in terms of probability", *Monthly Weather Review*, 78, 1-3.

Connoly, R.A. and Rendleman, Jr, R.J. (2008). Skill, luck and streaky play on the PGA tour. *Journal of the American Statistical Association*, 103(481), 74-78.

Decroos, T., Bransen, L., Van Haaren, J. and Davis, J. (2018). "Actions speak louder than goals: Valuing player actions in soccer", *Proceedings of the 25th ACM Conference on Knowledge Discovery and Data Mining (KDD 2019)*, retrieved Sep 16/20 at https://people.cs.kuleuven.be/∼tom.decroos/reports/kdd19_tomd.pdf

Duckworth, F.C. and Lewis, A.J. (1998). A fair method of resetting targets in one-day cricket matches. *Journal of the Operational Research Society*, 49, 220-227.

Fernández, J., Bornn, L. and Cervonne, D. (2019). "Decomposing the immeasurable sport: a deep learning expected possession value framework for soccer", *MIT Sloan Analytics Conference*, retrieved Sep 16/20 at www.sloansportsconference.com/wp-content/uploads/2019/02/Decomposing-the-Immeasurable-Sport.pdf

Genuer, R. and Poggi, J-M. (2020). *Random Forests with R*, Springer: New York.

Gudmundsson, J. and Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys*, 50(2), Article 22.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News, The Newsletter of the R Project*, Vol 2/3, 18-22.

Macdonald, B. (2012). "An expected goals model for evaluating NHL teams and players", *MIT Sloan Analytics Conference*, retrieved Sep 16/20 at www.hockeyanalytics.com/Research_files/NHL-Expected-Goals-Brian-Macdonald.pdf

Pollard, R., Ensum, J. and Taylor, S. (2004). "Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space", *International Journal of Soccer and Science*, 2, 50-55.

Rudd, S. (2011). "A framework for tactical analysis and individual offensive production assessment in soccer using Markov chains", In *New England Symposium on Statistics in Sports*, retrieved Sep 16/20 at http://nessis.org/nessis11/rudd.pdf

Sarkar, S. and Kamath, S. (2022). Does luck play a role in the determination of the rank positions in football leagues? A study of Europe's "big five". *Annals of Operations Research*, To appear.

Swartz, T.B. (2017). "Research directions in cricket", *Handbook of Statistical Methods and Analyses in Sports*, Editors Albert, J.A., Glickman, M.E., Swartz, T.B. and Koning, R.H., Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton, 445-460.

Weissbock, J. (2014). Forecasting success in the National Hockey League using in-game statistics and textual data. *PhD dissertation, Université d'Ottawa.*

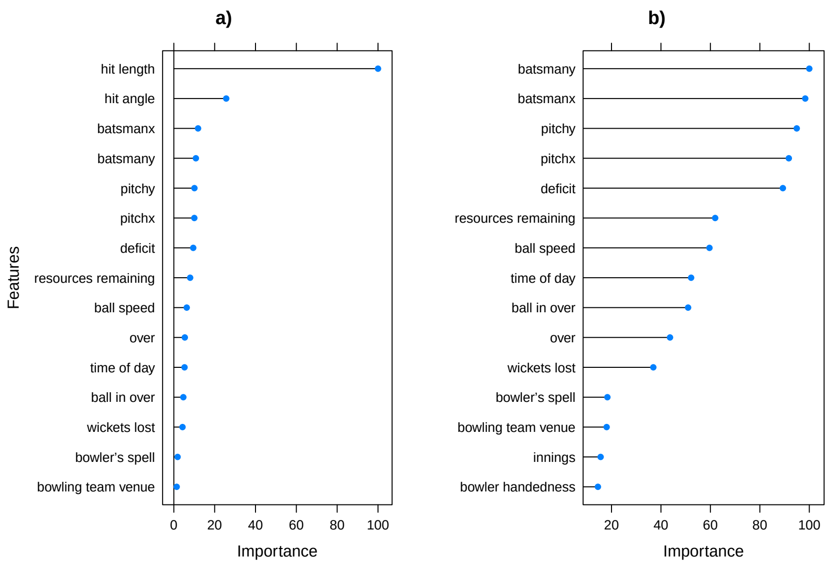

Figure 1: Importance plots of the features for the men's game corresponding to Analysis A and Analysis B.

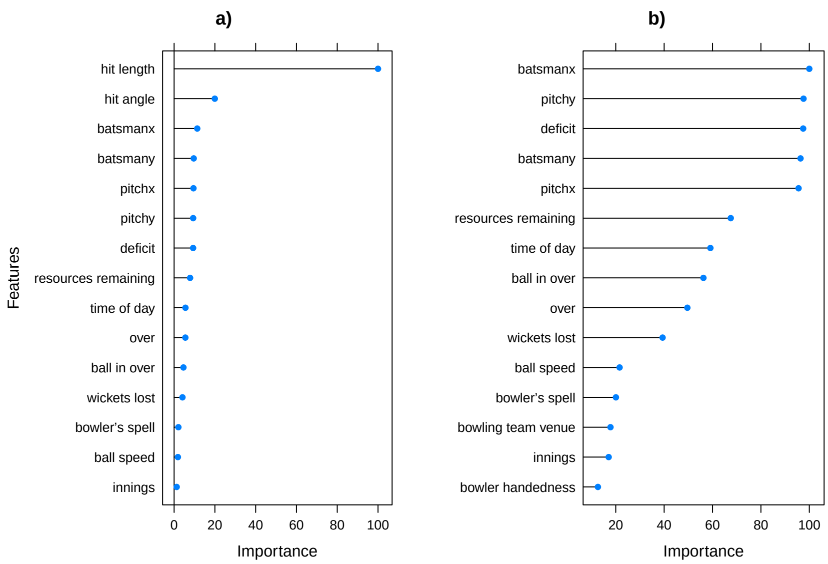

Figure 2: Importance plots of the features for the women's game corresponding to Analysis A and Analysis B.

Feature (Data Details)	Bowling Variable	Batting Variable	Percent Missing Observations
bowling team venue (home, neutral, road)	Y	N	0.00
innings (1, 2)	Y	N	0.00
time of day (hour using 24 hour clock)	Y	N	0.00
batsman handedness (L, R)	Y	N	0.00
bowler handedness (L, R)	Y	N	0.00
bowler's style (pace, spin)	Y	N	0.00
bowler's spell (1, 2, 3, 4)	Y	N	0.00
bowler's speed (fast, medium, slow)	Y	N	0.00
powerplay (Y, N)	Y	N	0.00
over (1, ..., 20)	Y	N	0.00
ball in over (1, ..., 6)	Y	N	0.00
wickets lost (0, ..., 9)	Y	N	0.00
resources remaining (0.0 to 100.0)	Y	N	0.00
deficit (integer)	Y	N	0.00
ball speed (continuous in km/hr)	Y	N	69%
ball rpm (continuous)	Y	N	99%
pitchx (horizontal ball landing rel to wicket)	Y	N	0.00
pitchy (vertical ball landing rel to wicket)	Y	N	0.00
batsmanx (horizontal ball landing rel to batsman)	Y	N	0.00
batsmany (vertical ball landing rel to batsman)	Y	N	0.00
hit angle (angle ball hit by batsman)	N	Y	0.00
hit length (distance ball stopped after hit)	N	Y	0.00
temperature (cold, moderate, hot)	Y	N	2%
humidity (dry, moderate, humid)	Y	N	2%
cloud cover (light, medium, heavy)	Y	N	2%

Table 1: Selected features (covariates) and related information used in the estimation of batting outcomes.

Prediction	Actual Outcome							
	Wicket	0	1	2	3	4	5	6
Wicket	1232	123	44	23	0	6	0	0
0	1377	27361	5483	83	4	4	11	0
1	2368	3873	31693	7182	410	305	4	12
2	146	3	543	1042	265	102	0	3
3	1	0	1	8	10	0	0	0
4	44	0	112	195	100	9923	3	96
5	0	0	0	0	0	0	0	0
6	35	1	66	79	15	44	1	4015

Table 2: Confusion matrix corresponding to Analysis A (bowling and batting features) for men based on a validation set of 98,450 observations.

Prediction	Actual Outcome							
	Wicket	0	1	2	3	4	5	6
Wicket	124	100	183	64	2	34	0	31
0	1697	16007	9594	2079	344	4174	10	885
1	3313	14820	27714	6294	429	5693	9	3056
2	31	51	121	65	4	38	0	29
3	0	0	1	0	0	1	0	0
4	30	375	309	100	25	432	0	104
5	0	0	0	1	0	0	0	0
6	8	8	20	9	0	12	0	21

Table 3: Confusion matrix corresponding to Analysis B (bowling features only) for men based on a validation set of 98,450 observations.

Prediction	Actual Outcome							
	Wicket	0	1	2	3	4	5	6
Wicket	1610	170	32	8	0	0	0	0
0	2683	46023	9924	162	11	4	11	1
1	2267	4687	38943	7879	412	4	7	7
2	143	7	915	1772	324	34	0	2
3	0	0	6	24	26	0	0	0
4	5	0	33	28	10	13564	1	70
5	0	0	0	0	0	0	0	0
6	9	1	40	32	9	41	0	1574

Table 4: Confusion matrix corresponding to Analysis A (bowling and batting features) for women based on a validation set of 133,515 observations.

Prediction	Actual Outcome							
	Wicket	0	1	2	3	4	5	6
Wicket	34	49	61	4	1	10	1	34
0	2765	31533	18581	3576	373	6598	11	2765
1	3897	19141	31017	6255	414	6865	7	3897
2	14	44	71	19	3	26	0	14
3	0	2	1	1	0	1	0	0
4	7	118	158	47	1	144	1	7
5	0	0	0	0	0	0	0	0
6	0	1	4	3	0	3	0	0

Table 5: Confusion matrix corresponding to Analysis B (bowling features only) for women based on a validation set of 133,515 observations.

Forecast	p(wicket)	$p(0)$	$p(1)$	$p(2)$	$p(3)$	$p(4)$	$p(5)$	$p(6)$
Uniform Discrete	0.166	0.166	0.166	0.166	0.000	0.166	0.000	0.166
T20 Proportions	0.056	0.305	0.404	0.071	0.000	0.112	0.000	0.048

Table 6: Probability estimates associated with two competing forecasts.

Bowler	Country	Balls Bowled in Validation Set	Career Economy Rate	xER (Analysis A)	xER (Analysis B)
Steketee, Mark	Australia	595	8.90	8.37	7.53
Lyon, Nathan Michael	Australia	620	7.21	7.28	7.74
Abbott, Sean	Australia	1185	8.54	8.54	7.82
Hauritz, Nathan	Australia	533	7.56	7.75	7.79
Archer, Jofra	England	512	7.65	7.43	7.70
Maxwell, Glenn	Australia	802	7.71	7.68	7.36
Ahmed, Fawad	Australia	933	6.96	6.88	8.13
Zampa, Adam	Australia	1306	7.29	6.76	7.94
McKay, Clinton	Australia	950	8.07	8.41	7.46
Boyce, Cameron	Australia	1109	7.65	7.72	7.96

Table 7: The expected economy rate statistic xER for 10 randomly selected men bowlers.

Bowler	Country	Balls Bowled in Validation Set	Career Economy Rate	xER (Analysis A)	xER (Analysis B)
Pike, Kirsten	Australia	708	7.34	5.93	5.67
Hepburn, Brooke	Australia	1454	6.97	6.70	6.15
King, Emma	Australia	1379	6.41	6.06	6.30
Birkett, Haidee	Australia	553	6.75	6.67	6.28
Kearney, Emma	Australia	657	6.22	6.33	5.73
Chappell, Renee	Australia	634	8.28	5.47	6.45
Elwiss, Georgia	England	604	5.92	6.90	6.55
Biss, Emma	Australia	521	5.22	6.01	6.08
Elliott, Sarah	Australia	921	6.20	5.72	6.24
Coyte, Sarah	Australia	2521	6.10	6.25	6.12

Table 8: The expected economy rate statistic xER for 10 randomly selected women bowlers.