Predicting International Success of Pace Bowlers in T20 Cricket

Ali Iltaf*, Richard Allmendinger**, Ali Hassanzadeh** and Richard Kingston**

University of Manchester, Manchester, United Kingdom
ali.iltaf@student.manchester.ac.uk
richard.allmendinger@manchester.ac.uk
ali.h@manchester.ac.uk
richard.kingston@manchester.ac.uk

Abstract

This study investigates the extent to which domestic T20 performance metrics can predict international success for pace bowlers in cricket. Using ball-by-ball data from over a decade of domestic and international T20 matches provided by the England and Wales Cricket Board (ECB), we engineer a comprehensive set of player-level features, including ball-tracking variables and outcome-based statistics. Success at the international level is evaluated using a Net Contribution metric adapted from the Duckworth-Lewis methodology. To identify key predictors, we apply feature selection techniques such as minimum redundancy maximum relevance (mRMR) and correlation clustering. Several regression models, including Random Forest and XGBoost, are trained and evaluated, with Random Forest achieving the best performance ($R^2 = 0.53$). Model interpretation using SHAP values reveals that a bowler's boundary percentage, dot ball percentage and percentage of their wickets taken that were caught are among the most influential features. These findings offer data-driven insights for selectors and talent scouts seeking to identify and fast-track promising pace bowlers from domestic leagues.

1 Introduction

T20 cricket is a short format of cricket designed to produce fast-paced games with an emphasis on scoring runs quickly. The England and Wales Cricket Board (ECB) manages the international cricket team as well as the domestic leagues, and it will always remain in the interest of the ECB to be able to identify new talent for the international team. The dynamics of the game at the T20 level do not translate perfectly to the international level, and it has been the case that players that have performed well in domestic leagues cannot uphold the same level of performance at the international stage.

This paper aims to aid this decision making process by asking: Can domestic T20 performance metrics predict a pace bowler's success in T20 Internationals? Performance is evaluated using a wide array of metrics which use ball-tracking statistics and match events to provide measures of bowlers' bowling ability and the outcomes of their bowling. Success is measured by the Average Net Contribution, which is the average difference in runs conceded and the expected number of runs scored over all deliveries bowled by a specific bowler in a match.

2 Background

There are several studies that attempt to predict player performance based on previous performance. Most of these studies use traditional metrics including strike rate and runs scored for batsmen and economy and wickets taken for bowlers. In order to differentiate players, it may also be required to go further into the details of player performance.

Asad et al. (2022) use commentary data to determine how many balls a batsman left, missed and hit to calculate the control of a batsman. This control measure was then used to calculate the 'Effective Runs', which is a metric proposed in the paper. Rupai et al. (2020) use pitch and weather data alongside ball tracking data to predict the outcome of each ball in a match. Mody et al. (2021) propose a formula for batting form, a pressure index and account for which team is the opposition. The study was in the context of the Indian Premier League (IPL) so the opposition team feature was a categorical variable which indicated one of the 8 IPL teams, but the opposition team cannot always be used as a factor if the teams are unknown. Mody et al. (2021) also uses classification to group players into scoring bands. The players in the highest rank are predicted to score the most runs. The problem with grouping players like this is that players of a different caliber can be grouped together in the same band, and it is made more difficult to make a comparative judgement of similarly ranked players.

A major factor in sports performance prediction is deciding what to use as a performance evaluation metric. Studies commonly used runs scored for a batsman or economy or wickets taken for a bowler as the target variable. The issue that arises with only using those metrics to profile performance is that it does not take into account the wholistic performance of the player. Lewis (2005) proposes two context-aware metrics called the Net Contribution and Resource Average. These metrics take into account the amount of wickets and overs remaining at each stage of the game and base the player's score on the aggregation of these resource contributions over every delivery. Lemmer (2002) proposes the Combined Bowling Rate (CBR), which is the harmonic mean of bowling strike rate (balls per wicket), economy rate (runs per over) and the runs per wicket, but this metric can only be applied to bowlers and is derived directly from other metrics which would be used as features. Thomson et al. (2021) propose a contextual batting score to measure batting and bowling performance when a team is batting second, but this metric can only be used to measure performance in the second innings.

Another aspect that is not seen in the literature is an interpretation of the models. In order to draw conclusions from the models to help inform decision-making, it is important that the driving factors for prediction for each model are considered. By making sure that the models are interpretable, it can also be checked that the models are making sense, and the relationship between the predictors and outcome have the desired relationship. For example, it would not make sense for a bowler's performance rating to be positively correlated with the bowler's economy (runs conceded per over).

3 Data

The data used for this study was provided by the ECB. This data includes ball-by-ball data on all professional T20 matches, including both international and domestic matches, from the start of 2010 up until 23rd October 2024. The data includes key details from the match the delivery was played in and more detailed data on each delivery, such as the information that can be found about the scorecard, shot and delivery types, foot

movement for the batsman, as well as some ball-tracking data.

Before analysis, the raw ball-by-ball match data was transformed to a player-level format, with rows representing individual players and columns capturing various performance metrics. The dataset was first divided into international and domestic subsets, with the former used for training and testing, and the latter reserved for prediction. To account for changes in performance across a player's career, statistics were further aggregated into age groups: 18–24, 25–28, 29–32, 33–36, and 37–42. The first and last groups span wider age ranges to ensure a sufficient number of matches for reliable statistics, minimizing distortion from outlier performances. After this aggregation, player records with missing values were removed, resulting in a final dataset of 630 players across 141 features.

4 Methods

4.1 Player Performance Evaluation

The Net Contribution metric in cricket evaluates a player's impact by measuring the difference between actual runs scored or conceded and the expected runs based on the Duckworth/Lewis (D/L) model, calculated on a ball-by-ball basis (Lewis, 2005). It incorporates match context—specifically, overs remaining and wickets lost—offering a more situational and comprehensive assessment of performance than traditional metrics. By integrating both run rate and wicket impact into a single value, it avoids inflation from performances against weaker opponents and remains calculable in all scenarios, unlike metrics such as bowling strike rate or CBR. Due to the proprietary nature of the original D/L formula (Duckworth and Lewis, 1998), this research uses a modified version proposed by McHale and Asif (2013). Overall, the Net Contribution metric offers a more nuanced understanding of player performance compared to traditional aggregate statistics (Lewis, 2008).

4.2 Feature Selection

With the feature set and target variable prepared, machine learning models can now be trained on the data. Given the presence of 141 features and high multicollinearity, feature selection is essential to retain predictive power while improving interpretability. Instead of using dimensionality reduction methods like PCA or t-SNE—which transform features into uninterpretable combinations—this study employs Minimum Redundancy Maximum Relevance (mRMR) and correlation clustering. These methods maintain the original features' meaning, which is crucial for understanding the relationship between features and performance. mRMR selects features that are highly relevant to the target while minimizing redundancy (Peng et al., 2005), improving efficiency without iterative model retraining. Correlation clustering uses hierarchical clustering with Spearman correlations and Ward's linkage to group redundant features, from which a single representative feature is selected per cluster for model training.

4.3 Model Training

Using the selected features, regression models were trained to predict Average Net Contribution, with five models evaluated: Linear Regression, Support Vector Regression, Decision Trees, Random Forests, and XGBoost. These models represent a range of techniques from simple linear to complex ensemble and

kernel-based approaches, allowing for a balanced comparison across different data characteristics. Data was normalized before training, and model performance was assessed using Mean Squared Error (MSE) and R^2. The best-performing model was then applied to domestic player data to predict performance and rank pace bowlers accordingly.

5 Results

Figure 1 shows that among the evaluated models, Random Forest achieved the best predictive performance, followed by XGBoost and then Linear Regression. Random Forest's robustness and ability to generalize well without intensive hyperparameter tuning made it particularly effective, especially given the noisy nature of the data. Although XGBoost is a powerful model, its performance may have been hindered by its sensitivity to hyperparameter settings, making it more prone to overfitting without careful tuning. Linear Regression performed reasonably well, likely due to the presence of some linear relationships in the data, while Decision Trees and SVR underperformed due to overfitting and sensitivity to noise or suboptimal parameters. Among the feature selection methods, correlation clustering performed poorly as it often grouped and excluded key predictors, resulting in uninformative feature sets. The mRMR methods (MID and MIQ) performed similarly across models, with MID working better for XGBoost and Decision Trees, and MIQ better for Linear Regression and SVR. For Random Forest, both mRMR methods produced nearly identical results, with MIQ slightly ahead. While omitting feature selection produced the best raw performance in most cases, the resulting models lacked interpretability, with many features showing zero or negative importance. This justified the use of interpretable feature selection techniques like mRMR.

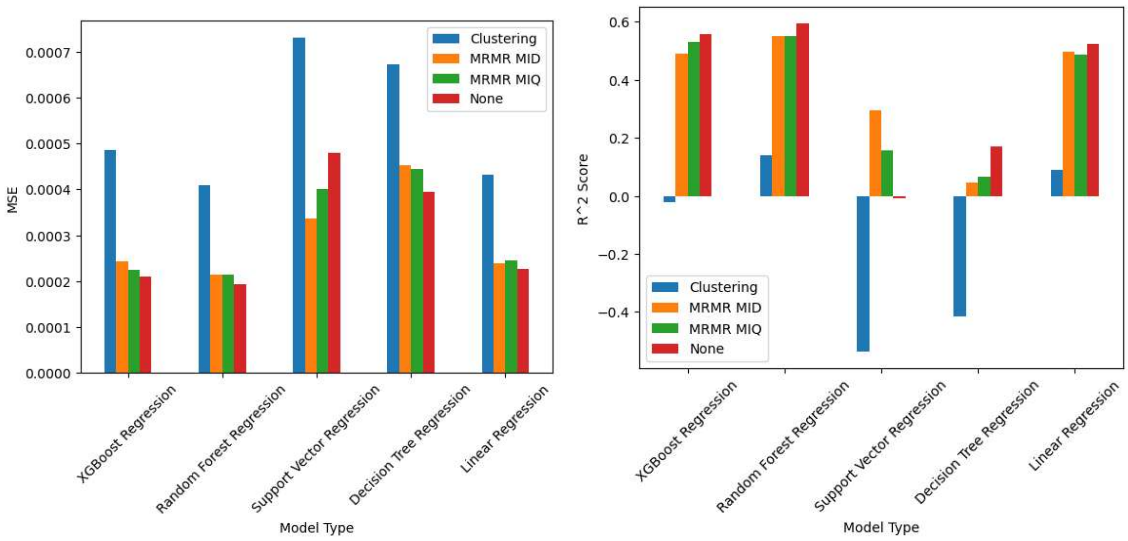

Figure 1: Mean Squared Error (left) and R^2 score (right) of each model type with each subset of features.

We select the Random Forest model using mRMR MIQ feature selection and inspect it more closely. Figure **??** shows a plot of the permutation importances of the features used in the model and Figure **??** shows a SHAP beeswarm plot, which ranks the features by their SHAP score and shows the relationship

between the feature and output for each sample. Both methods show that the three most important features are the Boundary %, the Dot Ball % and the % of wickets taken by the bowler that were caught. In both plots, the importance of these three features is significantly higher than the rest.

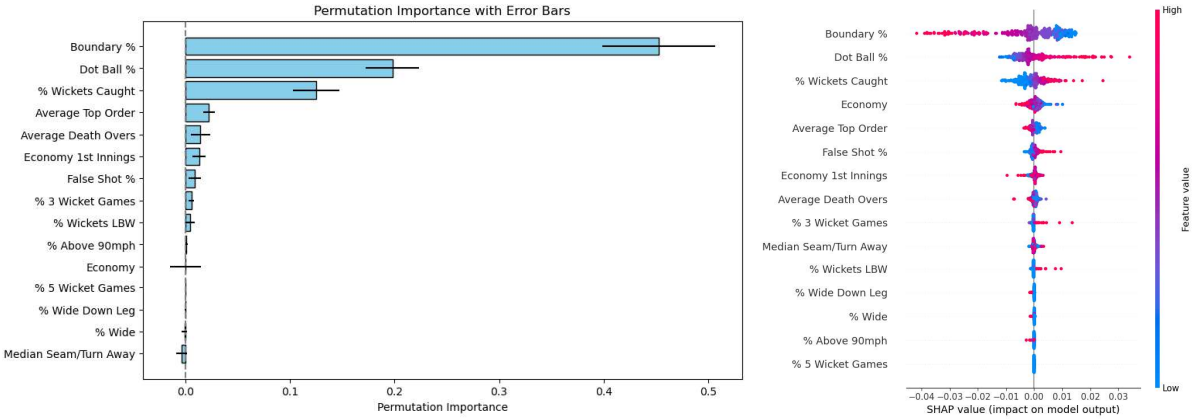

Figure 2: Permutation Importance plot (left) and SHAP beeswarm plot (right) of features used in the Random Forest model using mRMR MIQ feature selection.

Finally, using the MIQ Random Forest model, we predict the ouptut on the domestic data from the T20 Blast since 2016. Older tournaments are not included since we are interested in scouting recent performances. After removing retired players, we sort the players according to their predicted contribution, keeping only the most recent age group entry. The predicted top 10 bowlers for the England Cricket Team based on domestic performance are, in order of rank, Craig Overton, David Payne, David Willey, Benny Howell, Pat Brown, Tom Taylor, Jofra Archer, Matthew Waite, Paul Walter and Luke Fletcher.

6 Discussion

Out of the top 10 predicted pace bowlers, the fact that bowlers who have already made the international team, like Jofra Archer, Craig Overton and David Willey, shows that this method can predict which players are high performers. Other players who have played internationally have not played many games in the English domestic league, so they are not present.

The models show that there is a heavy emphasis on keeping the number of runs low, which is intuitive since T20 is a format that prioritises getting a large amount of runs quickly, rather than emphasising protecting the batter's wicket. Common knowledge suggests that the ability to bowl at high speeds and seam/swing bowling are crucial to breaking through to the international level. The models here show that these factors are not as important. This may be due to the fact that some metrics measure the bowlers actions, such as the line, length, and speed, whilst others measure the outcome of the delivery, such as the economy, strike rate, and boundary percentage. Further research should be undertaken on the causal relationship between these different types of metrics.

Acknowledgments

The data used for this research was provided by the England and Wales Cricket Board.

References

[1] Ahmad Al Asad et al. (2022) *Impact of a Batter in ODI Cricket Implementing Regression Models from Match Commentary.* In: 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1–6. DOI: 10.1109/CSDE56538.2022.10089357.

[2] F. C. Duckworth and A. J. Lewis. (1998) *A Fair Method for Resetting the Target in Interrupted One-Day Cricket Matches.* In: The Journal of the Operational Research Society 49.3, Palgrave Macmillan Journals, pp. 220–227. DOI: 10.2307/3010471.

[3] H.H. Lemmer. (2002) *The combined bowling rate as a measure of bowling performance in cricket.* In: South African Journal for Research in Sport, Physical Education and Recreation 24.2, pp. 37–44. DOI: 10.4314/sajrs.v24i2.25839.

[4] A J Lewis. (2005) *Towards fairer measures of player performance in one-day cricket.* In: Journal of the Operational Research Society 56.7, pp. 804–815. DOI: 10.1057/palgrave.jors.

[5] Lewis, A.J. (2008) *Extending the range of player-performance measures in one-day cricket*, In: Journal of the Operational Research Society, 59(6), pp. 729–742. DOI: https://doi.org/10.1057/palgrave.jors.2602379.

[6] Ian G. McHale and Muhammad Asif. (2013) *A modified Duckworth–Lewis method for adjusting targets in interrupted limited overs cricket.* In: European Journal of Operational Research 225.2, pp. 353–362. DOI: 10.1016/j.ejor.2012.09.036.

[7] Khush Mody, D. Malathi, and J. D. Dorathi Jayaseeli. (2021) *An Artificial Neural Network Approach for Classifying Cricket Batsman's Performance by Adam Optimizer and Prediction by Derived Attributes.* In: 2021 Smart Technologies, Communication and Robotics (STCR), pp. 1–7. DOI: 10.1109/STCR51658.2021.9588836.

[8] Hanchuan Peng, Fuhui Long and Ding, C. (2005) *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy*, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), pp. 1226–1238. DOI: https://doi.org/10.1109/TPAMI.2005.159.

[9] Aneem-Al-Ahsan Rupai, Md. Saddam Hossain Mukta, and A. K. M. Najmul Islam. (2020) *Predicting Bowling Performance in Cricket from Publicly Available Data.* In: Proceedings of the International Conference on Computing Advancements., pp. 1–6. DOI: 10.1145/3377049.3377112.

[10] James Thomson, Harsha Perera, and Tim B. Swartz. (2021) *Contextual batting and bowling in limited overs cricket.* In: South African Statistical Journal 55.1, pp. 73–86. DOI: 10.37920/sasj.2021.55.1.6.