Stats & Stumps: Using Machine Learning to Predict T20I Matches with Player and Venue Data

Archith Sharma[1]

[1]Texas Academy of Mathematics and Science

Abstract

Cricket is gaining popularity worldwide rapidly, and at the front is the newest format of the game, Twenty20 Internationals (T20I), and big data. This project attempts to predict cricket match outcomes using player-level performance metrics and machine learning models. A dataset of 1,029 T20I matches was analyzed, with player-level features engineered from batting and bowling statistics such as runs, strike rate, boundaries, wickets, economy rate, and maiden overs. These features were normalized by ground-specific scoring rates to account for venue effects. Four modeling approaches were compared: a simple heuristic based on total player impact, logistic regression with regularization, random forests, and support vector machines (SVMs). All models were implemented in `R` using packages such as `glmnet` and `e1071`. Logistic regression achieved the highest test accuracy of 70.24%, balancing predictive performance with model interpretability. The final model was used to generate win probabilities for both past and unseen matches, including the 2024 T20 World Cup Final (India vs South Africa) and a March 2025 match between New Zealand and Pakistan. These predictions reflect tight contests and demonstrate the model's ability to quantify match uncertainty. Overall, the project illustrates how venue-adjusted, player-level impact metrics can enable robust and interpretable match outcome prediction and player comparison in international T20 cricket.

Keywords:

T20I, Cricket, Machine Learning, Player Impact, Cricket Analytics, Venue Adjustment, Regression

GitHub Code for Data Processing, Figure Generation and Analysis, Modeling:
https://github.com/ArchithSharma/CricketPredictions

Contents

1 Introduction

Cricket is a bat-and-ball sport that originated in England and has grown into one of the most popular sports worldwide, particularly in countries like India, Australia, South Africa, and Pakistan. Played between two teams of 11 players each, the game involves batting, bowling, and fielding, with the primary objective being to score more runs than the opposing team. Cricket is played in different formats, ranging from the traditional five-day Test matches to the fast-paced Twenty20 (T20) format.

T20 cricket, the shortest and most explosive format of the game, revolutionized the sport with its fast-paced action and entertainment value. Each team gets a maximum of 20 overs (120 balls) to bat, encouraging aggressive stroke play, quick scoring, and strategic bowling. Since its official introduction in 2003, T20 cricket has gained massive popularity, leading to global tournaments such as the ICC T20 World Cup and franchise-based leagues such as the Indian Premier League (IPL) and the Big Bash League (BBL). With its thrilling finishes, power-hitting, and emphasis on adaptability, T20 cricket has attracted a new generation of fans while maintaining the essence of the sport.

The rise of predictive analytics in cricket has closely coincided with the rapid growth of the T20 format. With T20 matches fast-paced and often decided by fine margins, teams, analysts, and betting markets have increasingly turned to data-driven approaches to gain a competitive edge. The shorter format requires quick decision making, making real-time data analysis and predictive modeling crucial to optimize team strategies, player selection, and match predictions.

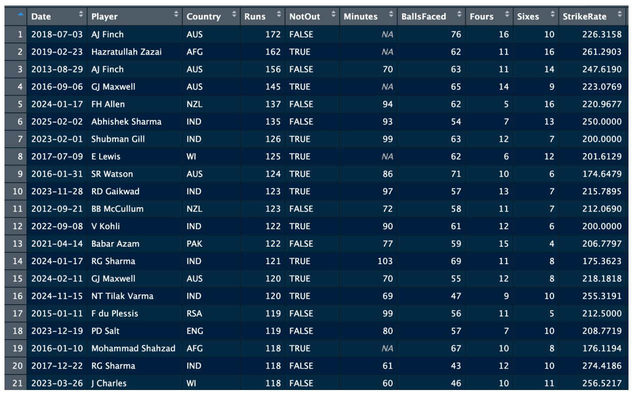
In this paper, a unique approach will be taken that focuses on a new impact factor (IF) of players for both batting and bowling, in addition to consideration of match venues. These factors will then be used to train 3 types of models commonly used in sports prediction: Regularized logistic regression, random forest classifier, and support vector machines (SVMs), with high test set accuracies of 67 - 70%. (Vistro et al., 2019)

The paper is organized as follows. Section 2 will specify the data used and the scope of the matches being considered and then explore the determination of the impact factor and the justification of the equations used for the IF. Section 3 focuses on analysis of the impact factor for different players and venues with rankings and case studies. Section 4 shows the determination of the heuristic, the total impact of the player, and how it will be used to predict matches. Section 5 will cover the modeling process and its results with an overall commentary on the impact factor and match predictions.

2 Materials and Methods

All data pre-processing, analysis and modeling was performed in R. (R Core Team, 2024) See Appendix A.1 to locate the code used on GitHub and regenerate the figures.

2.1 Description of Data



	Date	Player	Country	Runs	NotOut	Minutes	BallsFaced	Fours	Sixes	StrikeRate
1	2018-07-03	AJ Finch	AUS	172	FALSE	NA	76	16	10	226.3158
2	2019-02-23	Hazratullah Zazai	AFG	162	TRUE	NA	62	11	16	261.2903
3	2013-08-29	AJ Finch	AUS	156	FALSE	70	63	11	14	247.6190
4	2016-09-06	GJ Maxwell	AUS	145	TRUE	NA	65	14	9	223.0769
5	2024-01-17	FH Allen	NZL	137	FALSE	94	62	5	16	220.9677
6	2025-02-02	Abhishek Sharma	IND	135	FALSE	93	54	7	13	250.0000
7	2023-02-01	Shubman Gill	IND	126	TRUE	99	63	12	7	200.0000
8	2017-07-09	E Lewis	WI	125	TRUE	NA	62	6	12	201.6129
9	2016-01-31	SR Watson	AUS	124	TRUE	86	71	10	6	174.6479
10	2023-11-28	RD Gaikwad	IND	123	TRUE	97	57	13	7	215.7895
11	2012-09-21	BB McCullum	NZL	123	FALSE	72	58	11	7	212.0690
12	2022-09-08	V Kohli	IND	122	TRUE	90	61	12	6	200.0000
13	2021-04-14	Babar Azam	PAK	122	FALSE	77	59	15	4	206.7797
14	2024-01-17	RG Sharma	IND	121	TRUE	103	69	11	8	175.3623
15	2024-02-11	GJ Maxwell	AUS	120	TRUE	70	55	12	8	218.1818
16	2024-11-15	NT Tilak Varma	IND	120	TRUE	69	47	9	10	255.3191
17	2015-01-11	F du Plessis	RSA	119	FALSE	99	56	11	5	212.5000
18	2023-12-19	PD Salt	ENG	119	FALSE	80	57	7	10	208.7719
19	2016-01-10	Mohammad Shahzad	AFG	118	TRUE	NA	67	10	8	176.1194
20	2017-12-22	RG Sharma	IND	118	FALSE	61	43	12	10	274.4186
21	2023-03-26	J Charles	WI	118	FALSE	60	46	10	11	256.5217

Figure 1: Batting Data

The data is taken from ESPN Cricinfo, a reputable source containing in-depth statistics on all T20Is played and the players themselves. (ESPN, 2025a) To access the data in R, the cricketdata package made by Rob Hyndman (Hyndman et al., 2025) was used to access the Statsguru data in R as a dataframe. A screenshot of the batting dataframe is shown above in Figure 1.

The data is then filtered to only include T20Is between the test-playing full ICC members, defined in Appendix B. The data is also joined by match to determine the total impact factor of teams in their matches, both bowling and batting.

2.2 Derivation of Impact Factor

At the core of this analysis is the novel impact factor used in comparing batting and bowling impacts. To determine the formula for batting and bowling impact, correlation to the target metric between the total impact of both teams was prioritized in addition to similarity in the distribution of the impact factors each game.

The impact factor was manually tuned to a quantity that sufficiently addressed both criteria. They are defined below in Equations 1 and 2:

2.2.1 Definition of Bat and Bowl Impact

$$\text{Batting Impact} = (R \times 0.125) + ((SR - 130) \times 0.025) + (4s \times 0.3) + (6s \times 0.5) - \text{OutPenalty} \tag{1}$$

$$\begin{aligned} \text{where:}\quad & R = \text{Runs scored} \\ & SR = \text{Strike rate} = \left(\frac{\text{Runs}}{\text{Balls Faced}} \times 100 \right) \\ & 4s = \text{Number of fours hit} \\ & 6s = \text{Number of sixes hit} \\ & \text{OutPenalty} = \text{Fixed penalty of 0.5 if the batter is dismissed} \end{aligned}$$

Example:

A batter scores $R = 40$ runs off 28 balls, hits 3 fours, 2 sixes, and is dismissed.

$$SR = \frac{40}{28} \times 100 \approx 142.86, \quad \text{OutPenalty} = 0.5$$

$$\text{Batting Impact} = (40 \times 0.125) + ((142.86 - 130) \times 0.025) + (3 \times 0.3) + (2 \times 0.5) - 0.5$$

$$= 5 + (12.86 \times 0.025) + 0.9 + 1.0 - 2.5 = 5 + 0.3215 + 0.9 + 1.0 - 0.5 \approx \mathbf{6.72}$$

$$\text{Bowling Impact} = (W \times 3.25) + (M \times 3) - ((E - 7.4) \times 0.5) \tag{2}$$

$$\begin{aligned} \text{where:}\quad & W = \text{Number of wickets taken} \\ & M = \text{Number of maiden overs bowled} \\ & E = \text{Economy rate} = \left(\frac{\text{Runs Conceded}}{\text{Overs Bowled}} \right) \\ & 7.4 = \text{Baseline economy rate for adjustment} \\ & 0.5 = \text{Penalty per unit increase in economy above 7.4} \end{aligned}$$

Example:

A bowler takes $W = 2$ wickets and a $E = 6.85$ economy rate 3 and has no maiden overs.

$$\text{Bowling Impact} = (2 \times 3.25) + (0 * 3) - ((6.85 - 7.4) * 0.5)$$

$$= 6.5 + 0.55 * 0.5 \approx \mathbf{6.775}$$

Both batting and bowling equations consider a baseline adjustment for strike rate and economy rate based on the average score of around 149 in T20Is. (Statsguru, 2025) Both example players did very well in their respective games and had high similar impacts of around 6.75, showing the relevance of the heuristic.

Afterwards, the batting and bowling impacts were grouped by their totals by game for each team. Afterwards, the differences between total impact factors are used as a heuristic to be correlated to the winner of the match (interpreted numerically as 1, 0). As can be seen in the code, the correlation value is 0.773 for 1029 T20Is, and the R squared value is 0.598, approximately 0.6 which indicates high explainability.

2.2.2 ANOVA and Histogram Validation of Bat/Bowl Similarity

To validate the assumption that the distributions are similar, an analysis of variance (ANOVA) is performed by modeling the batting impacts on the bowling impacts and the result is displayed in Table 1. The p-value that is almost 0 shows that the distributions are very similar. In addition the visual validation of the histogram in Figure 2 shows how similar the total impact factors are, and that they can be used for modeling.

Term	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Total Bowling	1	19376	19376	95.95	$< 2 \times 10^{-16}$
Residuals	2068	417623	202		

Table 1: ANOVA Results

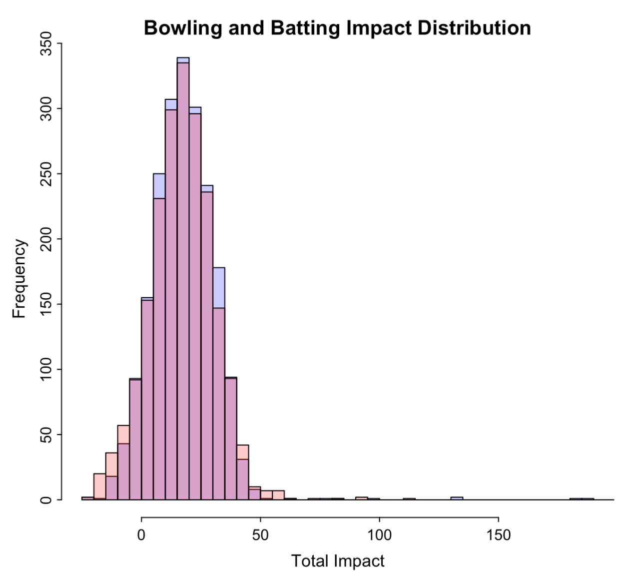


Figure 2: Impact Distributions

2.3 Determining Impact Factors to be used in Prediction

In this study, the performance of players in a T20 match is evaluated through a series of factors that encompass both their batting and bowling abilities. The impact of each player is adjusted for various game-specific conditions, including venue-based adjustments and player rankings. Below, the rationale behind the decision-making process is outlined for each factor used in the calculation of player impact.

2.3.1 Player Ratings and Impact Calculation

Each player's performance is assessed based on their batting and bowling ratings, which are derived from historical performance data. These ratings are extracted from the `player_rankings_2` dataset, which includes batting impact (`BatImpactperGame`) and bowling impact (`BowlImpactperGame`) for each player. These ratings reflect the overall contribution of a player to their team's performance in previous matches.

2.3.2 Ground Buffs (Venue Factors)

The performance of players is also adjusted based on the specific characteristics of the ground where the match is played. Each ground has unique attributes that can influence player performance, such as pitch conditions, weather, and altitude. The `venue_factors` dataset provides ground-specific adjustments known as "ground buffs". These buffs are applied to both batting and bowling ratings to account for the venue's impact on player performance. The following ground buffs are considered:

- **Batting Buff** (`BattingScale2`): This factor adjusts the batting impact based on ground-specific conditions such as pitch type, boundary size, and weather.
- **Bowling Buff** (`BowlingScale2`): This factor adjusts the bowling impact based on similar ground conditions that affect how bowlers perform, including pitch conditions and boundary dimensions.

2.3.3 Weighting by Player Ranking

The weight assigned to each player's performance is influenced by their relative ranking in the team. Batting and bowling impact are adjusted according to player rankings:

- **Batting Impact**: The top 8 batsmen (top and middle order) in the playing XI receive a weighted adjustment to their batting impact. The adjustment is proportional to the player's ranking, with higher-ranked players receiving a greater weighting. The weight is calculated as a function of the player's position in the batting order, with later batsmen receiving lower weights to reflect their reduced role in the match.

- **Bowling Impact**: Similarly, the top 6 bowlers are given weighted adjustments to their bowling impact, based on their ranking. The adjustment is calculated in a similar manner, with higher-ranked bowlers receiving greater weight. The bottom bowlers are excluded from the impact calculation, as they are less likely to bowl in the match.

2.3.4 Impact Per Game Adjustments

Once the initial ratings are adjusted for venue buffs and player rankings, the resulting impact values (`BatImpactperGame_2` and `BowlImpactperGame_2`) are calculated for each player. The calculations are performed iteratively for both teams, with batting and bowling impacts adjusted according to the above factors:

- For batsmen, the weight is scaled according to their position in the batting order.
- For bowlers, a similar scaling is applied, with the most prominent bowlers receiving higher weights.
- The adjusted values for each player are stored and used to update the overall team impact.

2.3.5 Calculation of Total Player Impact

Finally, the adjusted batting and bowling impacts are compiled into a total impact score for each player. This score is split into batting and bowling impacts, with the first 8 players in the batting order contributing to the batting total, and the next 6 players contributing to the bowling total. The impact scores for each player from both teams are then consolidated into a total impact matrix, which represents the overall performance potential of the two teams based on individual player ratings and match conditions.

In summary, the impact factors in this methodology are derived from a combination of player rankings, ground-specific adjustments, and positional weights. These factors provide a comprehensive assessment of player performance, accounting for both individual ability and contextual factors, such as venue and team composition.

2.4 Modeling

To predict T20I match outcomes based on player-level performance metrics, several classification models were implemented and evaluated using a dataset where each match was represented by the total impact scores of 28 players (14 per team). As a baseline, a simple heuristic predicted the team with the higher cumulative impact score as the winner. This naive method yielded an accuracy of **64%**, providing a benchmark for more advanced models.

A logistic regression model with elastic net regularization was trained using 5-fold cross-validation to tune the hyperparameters α and λ. (Tay et al., 2023) To convert predicted probabilities into binary outcomes, optimal classification thresholds were selected by averaging the best thresholds across ROC curves in 10-fold cross-validation. (Robin et al., 2011) Repeating this process over 30 iterations, the logistic regression model achieved a maximum accuracy of **70.24%**, outperforming the heuristic approach and all other models.

To enhance explainability, the coefficient weights were extracted from the logistic model. Features associated with top-order batters and key bowlers—such as the highest run scorers, highest strike rate performers, and wicket-takers—were consistently assigned large absolute coefficients, indicating their strong influence on match outcomes. Notably, impact scores from players in the top three batting positions had the greatest positive association with winning probability, while low-performing bowlers negatively influenced predicted outcomes. In addition, Batters 7-8 and Bowlers 5-6 had large coefficients, showing that the model respected lineup depth and all-rounder 'X-factor' performances.

A random forest model with 500 trees was also tried, tuning the `mtry` parameter via internal 5-fold cross-validation. This model achieved a maximum accuracy of **67.8%**. Feature importance rankings from the random forest agreed with the logistic regression model: early-order batters and strike bowlers had the greatest influence, though the non-linearity of the random forest also surfaced some interactions among lower-order players.

Lastly, a linear Support Vector Machine (SVM) was evaluated using a grid of cost values $C = 2^{-5}, \ldots, 2^{5}$ and 5-fold cross-validation. (Meyer et al., 2023) The best SVM model reached an accuracy of **69.7%**, comparable to the logistic regression model.

Among all approaches, logistic regression offered the best performance while also allowing for direct interpretability through model coefficients. This suggests that linearly weighted combinations of individual player impacts can be powerful predictors of T20I match outcomes.

3 Results

3.1 Player and Venue visualizations

Now that the impact factors are clearly defined, they can be used to construct rankings of players by batting, bowling, or all-around play and venue factors. Similarly, teams can be ranked by their recent impact factors. The full visualizations of the Top 15 in Career Impact and Impact per Game for batsmen are found in

Appendix A.2 containing Figures 4, 5, 6, and 7. All-rounders are defined as individuals in the top two-thirds of all players in batting and in the top 55th percentile of all players in bowling. The average venue impact was also min-max scaled to be used as a multiplier in the predictive modeling phase.

3.2 Model Results

The modeling results are organized below in Table 2.

Model	Mean Accuracy (%)	95% CI (%)
Heuristic (Total Impact Comparison)	64.00	[64.00, 64.00]
Logistic Regression (Elastic Net)	68.32	[66.42, 70.24]
Random Forest	66.08	[64.91, 67.80]
Support Vector Machine (Linear Kernel)	67.42	[65.53, 69.70]

Table 2: Model performance comparison based on mean accuracy and 95% confidence intervals over 30 iterations. Logistic regression showed the best performance on average and highest peak accuracy.

The following data in Table 3 is the model's output for select matches.

Table 3: Predicted Win Probabilities for Select T20I Matches

Match	Win Probability (%)
India vs South Africa (T20 WC Final)	India: 50.46
New Zealand vs Pakistan (Unseen Match, March 25)	New Zealand: 55.45
Bangladesh vs. West Indies Dec 19 2024	Bangladesh: 50.62
India vs. Pakistan Oct 23 2022	India: 64.5

4 Discussion

4.1 Analysis of Players and Venues

4.1.1 Player Analysis: Virat Kohli

Virat Kohli is considered by many to be the world's best batsman, scoring the third most runs of all time and having the second most total hundreds. (ESPN, 2025b) However, many have shown doubt about his performances in the shortest format of the game, especially due to concerns about his strike rate. (Desk, 2024)

The impact factor method considers him one of the best, if not the best T20I batsmen despite his strike rate. He is ranked the best batsman by total career batting impact, while he finds the third spot in batting impact per game. His low strike rate is not enough of a hindrance due to the constant amount of runs he amasses. The impact factor is able to find value in his ability to facilitate the innings.

4.1.2 Player Analysis: Ajantha Mendis

Ajantha Mendis (labeled BAW Mendis) was a spinner for Sri Lanka between 2008 and 2014. He is the only player to take 6 wickets in a T20I twice, and also won Emerging Player of the Year in 2008. Labelled a mystery spinner, when batsmen figured him out, he began to get hit all around the park. His biggest criticism was the lack of consistency, and he disappeared from Sri Lanka's XI in favor of Rangana Herath and other spinners. (Cricbuzz, 2025)

The impact factor method has a shortcoming when considering a spinner like Mendis. Recent matches do not get weighted more since the model does not have a time factor, and high performances in a small dataset could push his impact as high as the third rank. To address it in model predictions, adding a random effect to the player impact based on the number of games played could make a difference.

The theoretical framework would be that the more matches someone plays, the less deviation their impact score would have. Comparing Mendis and Kohli, where the former played one thirds the games the latter did, Mendis's variance parameter estimate would have 3 times the variance of Kohli's parameter and thus represent his inconsistency. Player dropping ensures the model stays representative regardless of consistency because inconsistent players tend to be dropped.

4.1.3 Player Analysis: Glenn Maxwell

Glenn Maxwell is considered the quintessential T20 player for the combination of his golden arm and power hitting. He can steal wickets from loose shots, and send loose deliveries 100 meters away. Most of his impact, however comes from batting. Unorthodox shots make him hard to stop when he gets going (ESPN, 2023), and his high batting impact scores and semi-decent bowling performances push him up the all-rounder list.

This is in contrast to someone like Wanindu Hasaranga higher on the list, who is a handy batsman lower in the order, but does most of his damage with the ball in his hands. Hardik Pandya is different from both, being middle-of-the-pack in bowling and batting. The impact factor method gives them high, yet roughly equal impacts which correspond to a correct aggregation.

4.1.4 Venue Analysis: England vs. Subcontinent Venues

The venue analysis by impact factor is mostly accurate to real-life ground impacts. 12 out of the top 20 best batting venues are located in either India or Pakistan, reflecting the popular opinion that many pitches in the subcontinent are batting paradises.

However, the grounds in England tend to be more balanced and favor the bowlers more. In total impact relative to batting, Lord's places 79th out of 105 venues. The famous slope at Lord's makes bowling relatively favorable as well. Southampton, The Oval and Headingley are all in the 60-70 range, reflecting England's moderate preference for bowling, showing that the scaling based on impact factor is an accurate measure of ground impacts.

4.1.5 Team Rankings Based on Recent Matches

Based on the impact factor, rankings based on the total accumulated impact can be constructed for the last 6 months. The data shows that India is far and away the most in-form team, followed by Australia and Afghanistan. South Africa and Pakistan are lower on the list, especially due to recent series losses to India and New Zealand, respectively. Ireland is also placed higher since they have not played the top teams in the last 6 months outside of Zimbabwe. However, overall, the list is an accurate representation of the current T20I landscape.

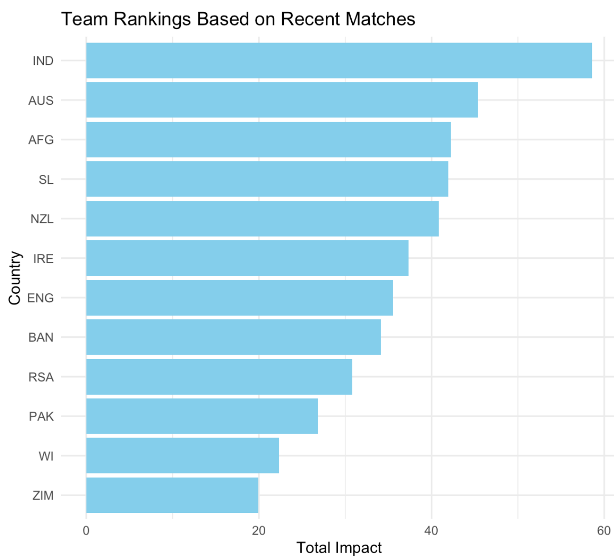


Figure 3: Rankings based on performance for last 6 months

4.2 Analysis of Models

Four different approaches were tested for predicting T20I match outcomes: a simple heuristic model, logistic regression with regularization, random forests, and support vector machines (SVM). The heuristic model, which selects the team with the highest total player impact, achieved an accuracy of approximately 64%, serving as a useful baseline. Logistic regression outperformed the others, reaching a maximum test accuracy of 70.24%, with hyperparameter tuning and threshold optimization via cross-validation. Random forests and SVMs followed closely with peak accuracies of 67.8% and 69.7%, respectively.

Logistic regression was selected for final prediction due to its strong performance and interpretability. The model provides probabilistic outputs, allowing uncertainty to be quantified. For example, in the 2024 T20 World Cup Final between India and South Africa, the model assigned India a win probability of 50.46%, reflecting an essentially even matchup which can be seen in the small result margin of 7 runs. Similarly, for the March 25 match between New Zealand and Pakistan, the model predicted a New Zealand win probability of 55.45%, indicating a slight edge.

These results suggest that logistic regression provides both competitive performance and actionable probabilities for match outcome predictions. Moreover, its explainability makes it suitable for understanding how individual player features contribute to team-level predictions.

5 Conclusions

This paper has been maximizing the value of a simple model in T20I prediction. More complex models can consider strategy with the toss before matches, as many pitches get better to bat on because of the dew in the night. (Krishnaswamy, 2022) In addition, recent player form can change the parameter passed into the logistic regression. For example, Virat Kohli was in a bad run of form during the T20 World Cup, only crossing 10 runs twice in the 7 matches before the final. Specific bowler-batsman combinations are not considered either, such as the Indian top order's issues with left arm pace bowlers.(Iyer, 2023) Overall, however, this paper's methods have had an accuracy of 70% in predictions and an explainable impact factor that can be used to quantify team performance. This framework can streamline the analysis of cricket and pit players and teams against each other objectively through the power of data.

6 Acknowledgements

I would like to thank the Wharton Sports Analytics and Business Initiative for the opportunity to share my work, and especially Professor Adi Wyner for orchestrating the Wharton High School Data Science competition that inspired my interest in sports analytics. I would also like to thank the cricket analytics community, especially ESPN Cricinfo and the R Project's open-source contributors for making data-driven sports analysis possible.

References

Cricbuzz. (2025). Ajantha mendis. *Cricbuzz*. Retrieved April 4, 2025, from https://www.cricbuzz.com/profiles/1390/ajantha-mendis#:~:text=Ajantha%20Mendis%20won%20the%20Emerging,has%20accomplished%20that%20feat%20twice.

Desk, T. S. (2024, April). 'usually he isn't this slow': Former india cricketers express concerns about virat kohli's strike rate. *The Times of India*. Retrieved April 4, 2025, from https://timesofindia.indiatimes.com/sports/cricket/ipl/top-stories/usually-he-isnt-this-slow-former-india-cricketers-express-concerns-about-virat-kohlis-strike-rate/articleshow/109615198.cms

ESPN. (2023, November). Afghanistan vs australia 07 11 2023. *ESPNcricinfo*. Retrieved April 4, 2025, from https://www.espncricinfo.com/series/icc-cricket-world-cup-2023-24-1367856/afghanistan-vs-australia-39th-match-1384430/live-cricket-score

ESPN. (2025a). Espncricinfo. *ESPNcricinfo*. Retrieved April 4, 2025, from https://www.espncricinfo.com/

ESPN. (2025b). Virat kohli records. *ESPNcricinfo*. Retrieved April 4, 2025, from https://www.espncricinfo.com/cricketers/virat-kohli-253802/tests-odi-t20-records

Hyndman, R., Gray, C., Gupta, S., Hyndman, T., Rafique, H., & Tran, J. (2025). *Cricketdata: International cricket data* [R package version 0.3.0, https://github.com/robjhyndman/cricketdata]. https://pkg.robjhyndman.com/cricketdata/

Iyer, R. (2023, September). Why do indian batsmen struggle against left-arm pacers? rohit sharma and co.'s record against left-armers detailed. *Sportingnews.com*. Retrieved April 4, 2025, from https://www.sportingnews.com/in/cricket/news/why-do-indian-batsmen-struggle-against-left-arm-pacers-rohit-sharma-and-cos-record-against-left-armers-detailed/aomdbaj3kqbxdyimaxsefxrm

Krishnaswamy, K. (2022, April). The (un)dew advantage: What choice do teams batting first have? *ESPNcricinfo*. Retrieved April 4, 2025, from https://www.espncricinfo.com/story/ipl-2022-pbks-vs-kkr-the-undew-advantage-what-choice-do-teams-batting-first-have-1308698

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). *E1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien* [R package version 1.7-14]. https://CRAN.R-project.org/package=e1071

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). Proc: An open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, *12*, 77.

Statsguru. (2025). Total match results: T20 internationals. *Cricinfo*. Retrieved April 4, 2025, from https://stats.espncricinfo.com/ci/engine/stats/index.html?class=3;orderby=matches;orderbyad=reverse;template=results;type=aggregate

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, *106*(1), 1–31. https://doi.org/10.18637/jss.v106.i01

Vistro, D., Rasheed, F., & David, L. (2019). The cricket winner prediction with application of machine learning and data analytics. *International Journal of Scientific & Technology Research*. https://www.kclas.ac.in/wp-content/uploads/2021/01/The-Cricket-Winner-Prediction-With-Application-Of-Machine-Learning-And-Data-Analytics-.pdf

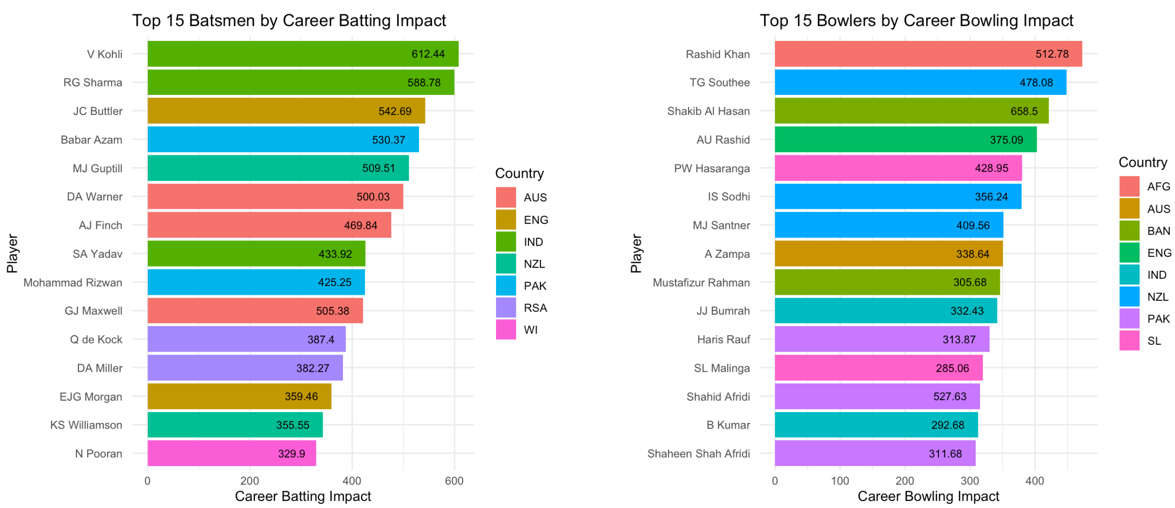
Appendices

A Code Used for Analysis, Player and Venue Visualizations

A.1 Code

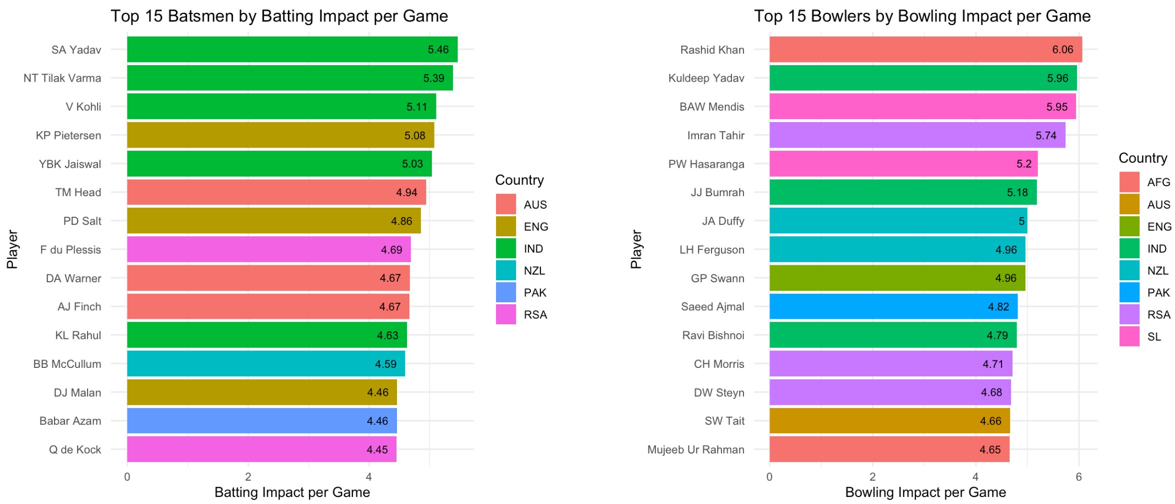
All code used to process data, generate figures, & tune parameters/validate model at https://github.com/ArchithSharma/CricketPredictions.

A.2 Player and Venue Diagrams



(a) Batting Career Impact (b) Bowling Career Impact

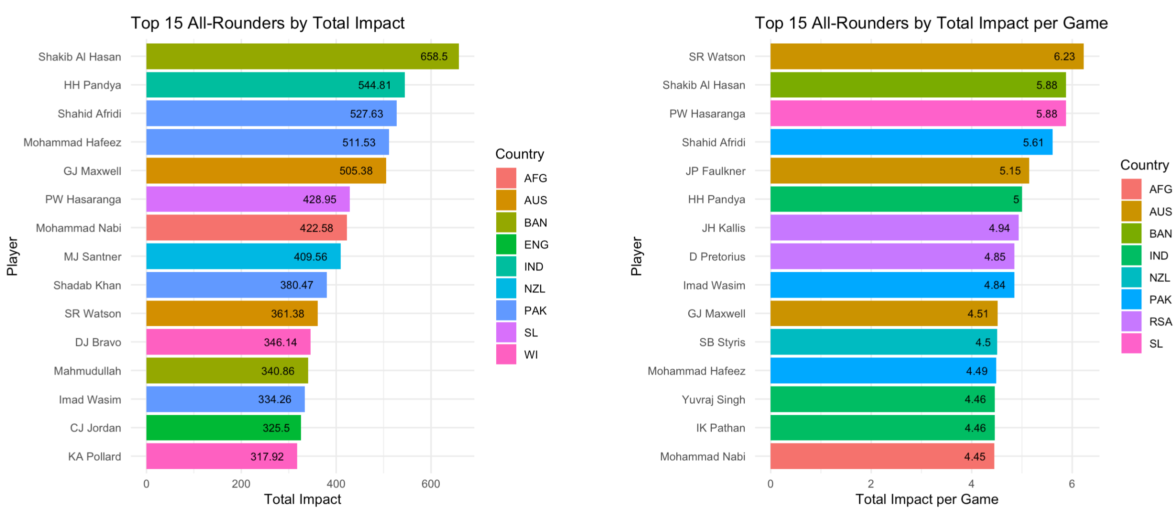
Figure 4: Total Career Bat/Bowl Impact Top 15



(a) Batting Impact by game

(b) Bowling Impact by game

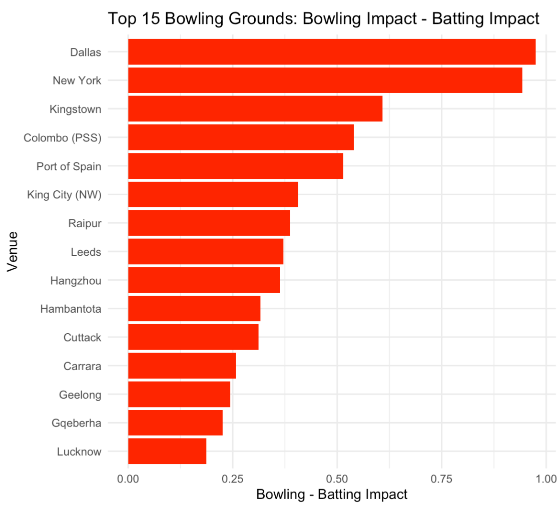
Figure 5: Bat/Bowl Impact Top 15 per game



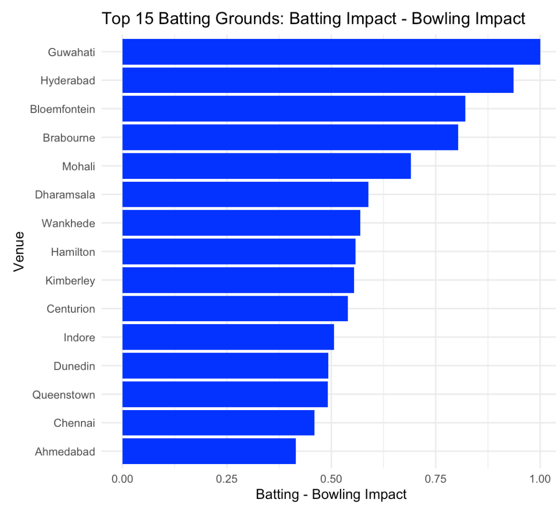
(a) All Rounder Total Impact

(b) All Rounder Impact by game

Figure 6: All Rounder by game/career impact



(a) Best Bowling Venues



(b) Best Batting Venues

Figure 7: Best Venues for Batting/Bowling

B Glossary of Cricket Terms Used

Batting Terms

Runs Total number of runs scored by a batsman.

Strike Rate The average number of runs scored per 100 balls faced. A higher strike rate indicates more aggressive scoring.

Boundary A scoring shot that results in four (4) or six (6) runs.

Fours & Sixes A four occurs when the ball crosses the boundary after touching the ground; a six is when it crosses without touching the ground.

Batting Impact Score A custom metric combining runs, strike rate, and boundaries to evaluate batting effectiveness.

Bowling Terms

Wickets The number of batsmen a bowler dismisses.

Maiden Overs Overs in which no runs are conceded.

Economy Rate Runs conceded per over bowled. A lower economy indicates more efficient bowling.

Spinner A bowler who bowls slow but gets the ball to turn, deceiving batsmen.

Pacer A bowler who bowls quick to try and beat batsmen with variations and leave them helpless.

Bowling Impact Score A weighted metric assessing bowling performance using wickets, maidens, and economy rate.

Match Context

T20I Twenty20 International – a short-format cricket match where each team bowls a maximum of 20 overs.

Venue Factor A numerical measure of how batting-friendly or bowling-friendly a stadium is, based on its average scoring.

Playing XI The eleven players selected to participate in a match.

Strategy Terms

All-Rounder A player who contributes significantly with both bat and ball.

Bowling Attack The group of bowlers in a team and their combined effectiveness.

Top Order The group of batsmen who come out to bat first and score most of the runs, typically the first three or four guys in.

Middle Order The group of batsmen that are typically all-rounders and power hitters to try and get quick runs near the end.

Test playing nations list: Afghanistan, Australia, Bangladesh, England, India, Ireland, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies, Zimbabwe.