Human Performance - Short Report

T20 Player Performance Index

Vibha Naiknavare

Carlo Lopez Hernandez

University of Notre Dame - Mendoza School of Business

MSBA - Sports Analytics

MSSA-60530 Human Performance

2024 Spring

T20 in cricket is a fast and exciting version of the game, defined by its short format compared to test cricket. This has changed how we look at cricket strategies and how we judge player performance. Traditional ways of measuring a player's value don't always tell us everything we need to know in T20 cricket. This shows the importance of creating a new more detailed way to evaluate players that fits the special nature of T20 cricket better.

Research Objective:

The primary goal of this research is to create a player performance index specifically for T20 cricket. This index aims to quantify each player's overall impact on the game, taking into account their specific roles, such as Batter, Bowler, and Wicketkeeper Batter, and the context of their performances. Through this project, we will focus on analyzing how these roles influence match outcomes and player rankings within the T20 format.

Problem Framing:

The main challenge is determining which performance indicators truly show how effective a player is in their role in T20 cricket. We need to look closely at different parts of cricket performance, like how aggressive a batter is, how efficient a bowler is, how quick a fielder is, and their overall game understanding. Our goal is to bring all these aspects together into a single, clear index that reflects each player's role and contribution to the game.

Data Overview:

We gathered our data from Kaggle, only focusing on the T20 cricket tournament. We merged a total of 5 datasets to compile comprehensive statistics from T20 cricket matches. The datasets include the "bat_t20" dataset containing batting metrics such as runs scored and strike rates. The "bowl_t20" dataset covers bowling statistics, including wickets taken, bowling averages, and economy rates. The "all_round_t20" dataset is used to combine information for players who contribute in both batting and bowling, indicating their all-around performance. The "all_player" dataset included a roster of all players, including personal information, career stats, handedness, and role. Lastly, the "country" dataset holds records of country ID and name. This dataset was used to link a nation to a player. We merged these datasets using player and nation IDs to create

one large final dataset (further explained in the data cleaning section). Using this we could analyze player and team performances in the T20 cricket tournament (Appendix 1).

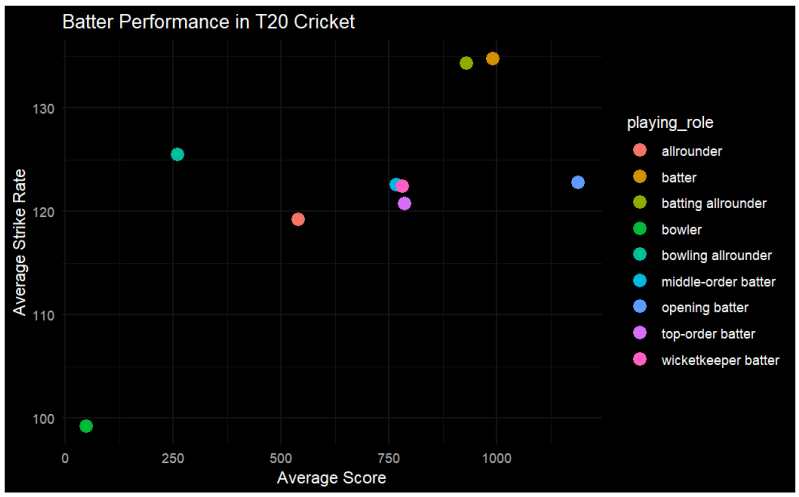
Exploratory analysis and visualizations

Batting Performance in T20 Cricket:

The scatter plot visualizes T20 cricket players' roles in terms of average scoring and strike rate, offering a view of their batting skills. Allrounders are showcased with varied average scores and strike rates, underlining their

versatility and dual contributions with bat and ball, making them integral to team dynamics. Batters, focused on run-scoring, often boast higher averages, with their strike rates reflecting a mix of aggressive and measured approaches. Batting Allrounders balance scoring with bowling capabilities, indicated by their respectable batting metrics. Bowlers, in contrast, typically have lower batting figures, as their main role is wicket-taking, though Bowling Allrounders demonstrate slightly enhanced batting prowess. Middle-order batters are pivotal in fortifying the innings, as their solid averages suggest while Opening Batters are distinguished by their aggressive strike rates, aiming for a robust start. Top-order batters, vital for scoring, exhibit stats that highlight their foundational role. Lastly, Wicketkeeper Batters display a spectrum of performances, acknowledging the varied demands of their position

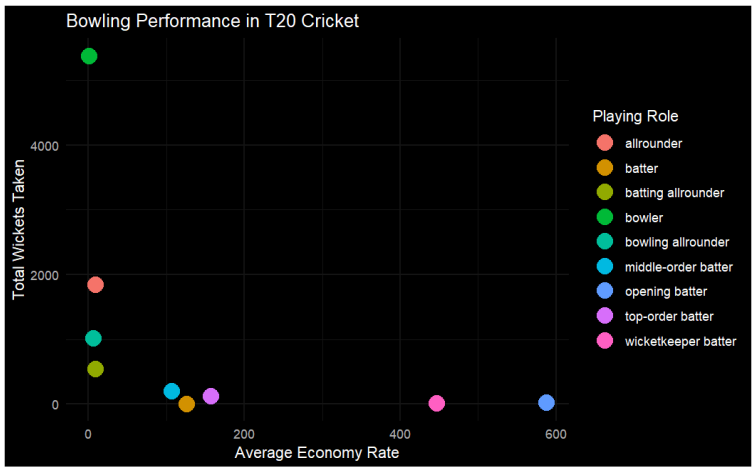
Figure 1: Batting Performance in T20



Bowling Performance in T20 Cricket: The scatter plot shows the bowling performance of T20 cricket players, with different colors representing their roles on the team. It plots how economical the players are against how many wickets they've taken: Middle-Order, Opening,

Top-Order, and Wicketkeeper Batters generally focus on run-scoring, which accounts for their lower wicket counts on the charts. Allrounders show a broad spectrum of economy rates, indicative of their varying degrees of bowling involvement, with some proving to be quite effective in wicket-taking. Batters, true to their primary role, rarely take wickets, as they seldom bowl. Batting all-rounders, while mainly run-scorers, do take wickets, reflecting their supplementary bowling skills. Bowlers, as expected, secure higher numbers of wickets, but their economy rates display a range, hinting at differences in bowling control. Bowling all-rounders balance wicket-taking with their ability to bat, yet they may not always bowl as economically as their specialist bowling counterparts.

Figure 2: Bowling performance in T20

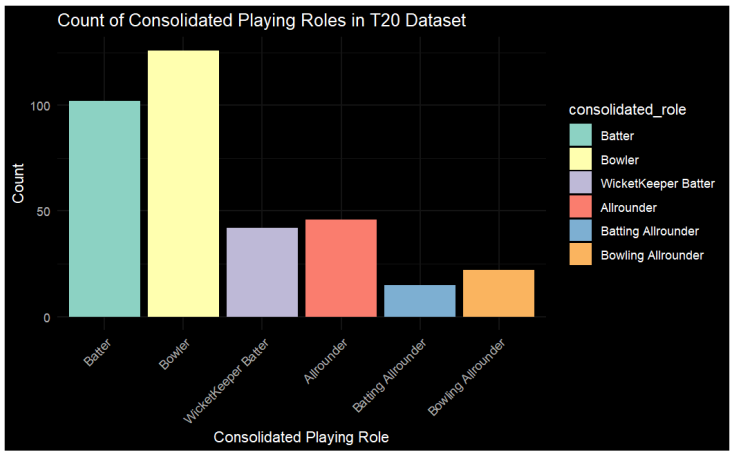


Data Cleaning:

Firstly, in the dataset, we refined the column names to make them clearer, especially after combining all the datasets. Columns that ended in .x, mostly bowling stats, were renamed to end with "_bowling" to indicate their focus on bowling. Similarly, columns ending in .y, which pertained to all-rounder performances, were renamed to "_allrounder". Later for further analysis, columns entirely missing data were removed to tidy up the dataset. The high_score column, which included non-numeric characters, was cleaned to retain only numeric values, allowing for straightforward numerical analysis. Furthermore, player roles were simplified into broad categories such as WicketKeeper Batter, Batting Allrounder, Bowling Allrounder, Allrounder, Batter, Bowler, and Others (see Figure 3). This step ensured the dataset was clean, organized, and primed for developing an effective Player Performance Index by focusing on the most relevant

information. Additionally, there are quite a few NA values in the dataset because not all batters are bowlers, so we replaced all the NA values with 0 to make the dataset more understandable.

Figure 3: Top 3 Consolidates playing roles in T20



Further Analysis:

Moving forward, we will be making clusters for each role- Batter, Bowler, and Wicket Keeper batters. By segmenting players based on their specific roles, we aim to gain deeper insights into the playing behaviors and performance characteristics unique to each role. Next, we will interpret the clustering results to understand the distinctive playing styles and strategies associated with each cluster. This will help us provide valuable insights into how players within each role contribute to their team's performance.

Furthermore, we will merge our cluster player dataset with T20 match results to perform predictive modeling, We will use logistic regression to predict match outcomes based on player clusters, relieving how different clusters correlate with results, and showing us the impact of player roles and behaviors on team performance. Finally, we will perform a machine learning model, XGBoost to enhance predictive model performance. By doing this we are aiming to give a full picture of how player roles in T20 cricket affect match results. By clustering players based on their roles, we'll see how different player groups impact the game. Also, we'll show how using predictive modeling can help cricket management and coaching make smarter decisions.

Scaling before Clustering:

In our data preprocessing, we scaled our metrics for all three roles. This step was important because players in roles face different situations in a match, for example number of balls faced or

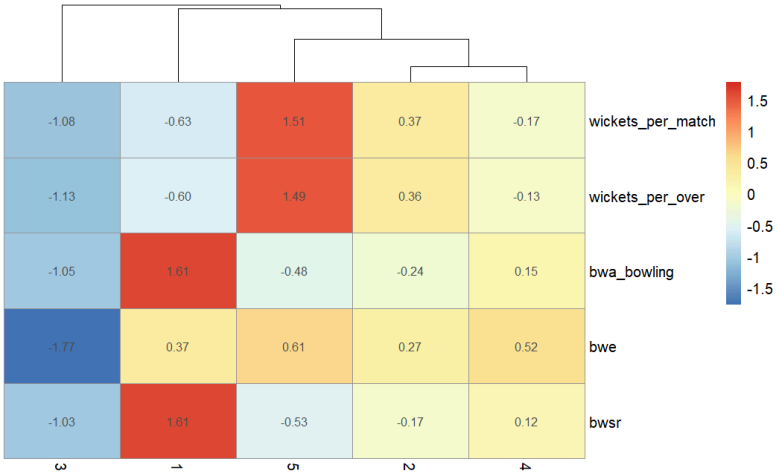
overs bowled. By scaling the metrics, we ensure that each player's performance is comparable regardless of these variations. For batters and wicketkeepers, we calculated rates like boundary rate and century rate based on their innings and balls faced. Similarly, for bowlers, we considered metrics like wickets per match and over, accounting for the number of matches played and overs bowled. This will help us analyze player performance more accurately and gain meaningful insights across different player roles.

Clustering and Modeling:

Bowlers Clustering :

We implemented k-means clustering on normalized bowling data to categorize players into distinct groups based on their performance metrics. By employing the silhouette method, we determined the most coherent and distinct number of clusters, which was 5. Each cluster was then analyzed and visualized through a heatmap of the centroids as seen in Figure 4, providing a clear depiction of the unique characteristics and performance profiles of each group of bowlers.

Figure 4: Clusters for bowling



Player performance according to cluster: Interpretation

Cluster 3 includes players with significantly below-average performances, likely ineffective in matches. Cluster 1 features bowlers with excellent averages and wicket-taking abilities, highlighting their effectiveness. Cluster 5 consists of aggressive bowlers who take wickets at a high cost, demonstrating risk. Cluster 2 comprises players who focus on maintaining low

economy rates and effective bowling. Cluster 4 contains bowlers who balance wicket-taking with economical performances.

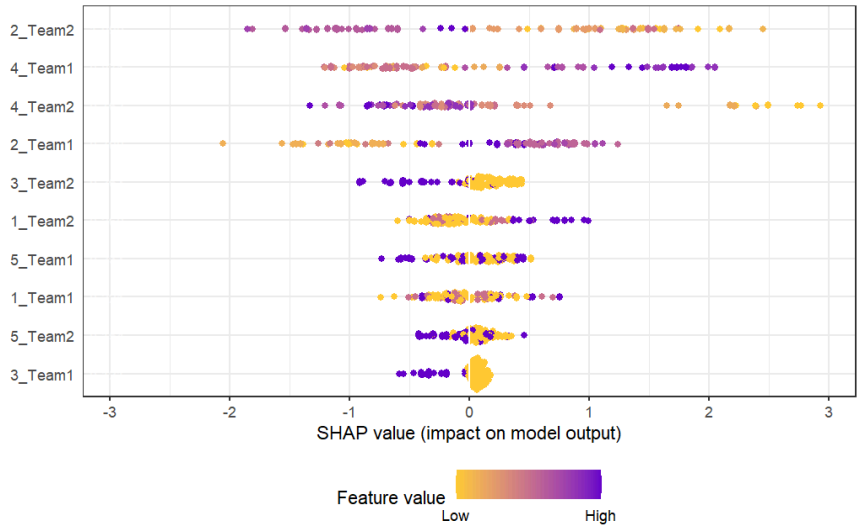
Logitic regression and XG boost:

In our analysis, we prepared the World Cup results data by focusing on the outcomes for Team 1, creating a binary variable to indicate whether Team 1 was the winner of each match. We utilized cluster-based features, representing performance metrics for both Team 1 and Team 2, to predict these match outcomes. This approach allowed us to model and predict the likelihood of Team 1 winning based on the relative strengths and strategic formations of the teams as described by their cluster assignments.

The logistic regression model achieved an accuracy of 66.1% with a balanced accuracy of 61.89%, indicating moderate effectiveness in classifying the outcomes for both classes. The sensitivity and specificity rates were 53.33% and 70.45% respectively, showing a stronger ability to correctly identify negatives (non-winners) than positives (winners). To make our model more accurate, we ran an XG boost model.

The XGBoost model achieved an accuracy of 66.1% before tuning, with a balanced accuracy of 61.89%, which shows a moderate level of effectiveness in correctly classifying both positive and negative outcomes. After hyperparameter tuning, the XGBoost model's performance was enhanced, exhibiting a balanced accuracy of 60.83% and improved model reliability and consistency, as indicated by a Kappa value of 0.2102

Shap Graph



Results:

Cluster 1 (Excellent Bowlers): These players are crucial in T20 formats for their consistent effectiveness in limiting scores and securing wickets, significantly enhancing the team's winning prospects in high-pressure games.

Cluster 5 (Aggressive Bowlers): Although risky due to higher run rates, their ability to shift game momentum with rapid wickets makes them valuable against formidable batting sides.

Cluster 2 (Cost-Effective Bowlers): Useful during the middle overs, these bowlers help control the game pace and manage run rates, though they are less impactful than the top performers.

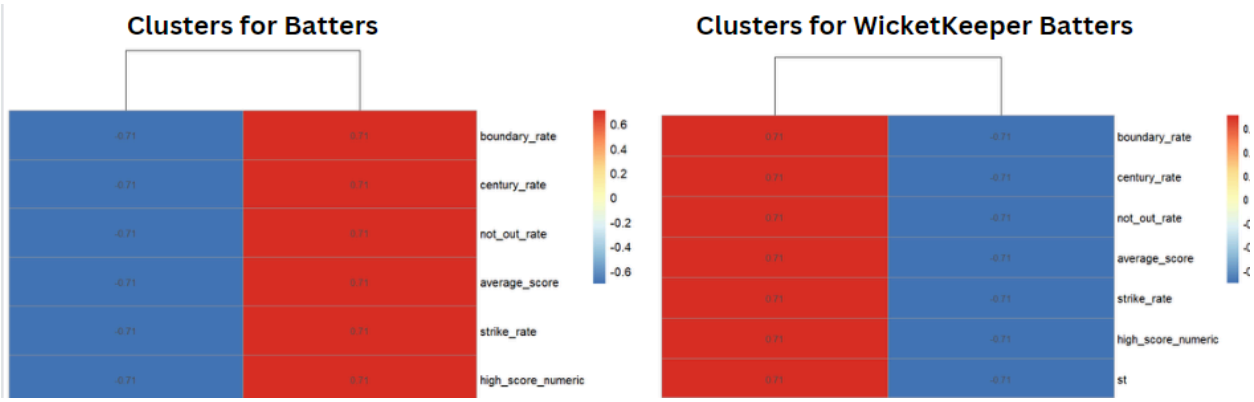
Cluster 4 (Balanced Bowlers): These bowlers offer versatility with a mixed impact on games; their balanced approach can sometimes be exploited by opposing teams. Cluster 3 (Below-Average Performers): Often a liability in the intense T20 World Cup environment, these player's roles should be limited or improved through additional training.

For a game like T20, controlling the scoring rate is critical due to the game's short duration. Therefore, Cluster 1 bowlers, known for their ability to effectively limit runs and secure crucial wickets, are essential for strategic decision-making in team management

Batters and WicketKeeping Batters:

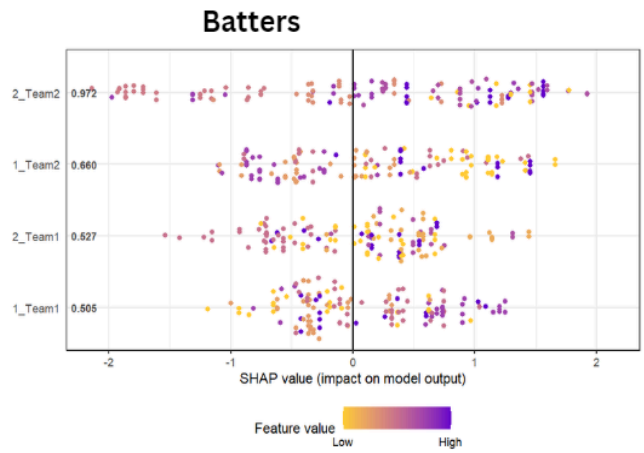
The cluster analysis (seen below) for batting and wicketkeepers yielded similar results. Both analyses showed values associated with the metrics of boundary rate, century rate, not_out_rate, average score, strike rate, and high_score_numeric. The reason for these similarities lies in the wicketkeepers and batters sharing distinct roles on the field—as wicketkeepers also bat. However, the variable "st" was added, signifying "stumps," which is a metric associated only with wicketkeeping, as players effectively get the opposing team's wickets and earn a point.

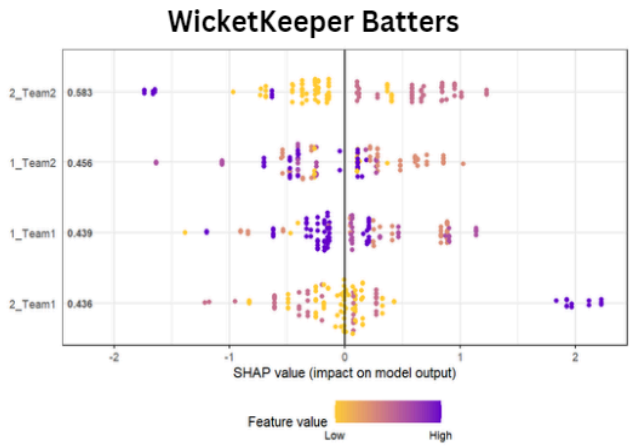
In the cluster analysis for batters, there was a strong positive correlation of 0.71 between these variables and Cluster 2, marked in red. This suggests that players in Cluster 2 could offer a competitive advantage, leading countries to prefer players from this cluster for their rosters. Conversely, for wicketkeepers, a strong positive correlation of 0.71 was observed in Cluster 1. This indicates that having players from the first cluster could also provide a competitive advantage in wicketkeeping roles.



To test the results of cluster analysis, an XGBoost model was developed (as previously mentioned) and the results were input into a SHAP plot. The cluster was denoted by the first number in the variable name. As we were predicting wins and losses based on clusters, we can see that Cluster 2 for batting had the highest influence on the model with a value of 0.972 and was most likely to lead to a win. This supported our previous analysis showing the importance of Cluster 2 in the model and how players in this cluster would provide an advantage to teams.

Similarly, to evaluate the results for the wicketkeeper role, we integrated the analysis with the XGBoost model and generated a similar SHAP plot. These findings differed from our initial expectations; here, Cluster 2 had greater influence over the model, contradicting our earlier analysis that highlighted Cluster 1 as more significant. However, with deeper examination, these results are not extraordinary. Wicketkeeping fundamentally shares many similarities with batting, with the primary distinction being the inclusion of the “Stumps” variable. This similarity likely explains why Cluster 2 still significantly influences the model, similar to the batting SHAP plot analysis. Furthermore, the wicketkeeping SHAP plot reveals that while Cluster 2 has the most influence, the difference from the subsequent variables (both in Cluster 1) is minimal. Cluster 2's influence score is 0.5, with the next two scores in the 0.4 range. Despite the difference, Cluster 1 continues to have a similar impact as Cluster 2 in this model.





Conclusion and future work:

This model developed a specialized player performance index for T20 cricket, categorizing players into clusters that reflect their roles and effectiveness. The analysis revealed that certain clusters, particularly those involving key bowlers in Cluster 1 and Cluster 2, identified as the most influential for batters and wicketkeepers, were found to significantly impact winning probabilities in matches. The use of logistic regression and XGBoost models demonstrated the predictive power of these clusters, confirming their importance in strategic decision-making within T20 matches

To further increase the predictive power and practicality of the player performance index developed for T20 cricket, future work could focus on integrating more detailed data such as player fitness, injury, environmental conditions, and psychological factors for a deeper analysis of performance influence. Moreover, expanding this model to other formats like ODIs and Test cricket would extend its applicability and impact across the sport, improving player selection and team management strategies.

Appendix

Appendix 1:

Dataset abbreviations:

Id: Player Identifier
span: Career Span (Years Active)
matches: Number of Matches Played
innings: Number of Innings Played
not_out: Times Not Out
runs: total Runs Scored
high_score: Highest Score in an Innings
average_score: Batting Average
ball_faced: Balls Faced by the Batsman
strike_rate: Batting Strike Rate
100s: Centuries Scored
50: Half-Centuries Scored
0s: Ducks (Dismissed Without Scoring)
4s: Fours Hit
6s: Sixes Hit
sp_bowling: Career Span (Years Active)
bbi_bowling: Best Bowling in Innings (Best Figures in a Single Innings)
bbm: The best bowling match by the player
bl: The total number of balls bowled by the player
bwa_bowling:The bowling average of the player
bwe: The economy rate of the player
bwsr: Bowling Strike Rate (Average Balls Bowled Per Wicket Taken)
cd: The number of times the player has achieved a 5-wicket haul
fw_bowling:The number of times the player has taken 4 wickets in an innings)
fwk: The number of times the player has taken a wicket with their first ball in an innings
in: The total number of innings bowled by the player
md: Maiden Overs Bowled
mt: The number of matches played by the player
Ov: The total number of overs bowled by the player
pr_bowling: The number of times the player has achieved a 3-wicket haul
tw: The total number of wickets taken by the player
wk_bowling: Wickets Taken (in Bowling)
bbad: Best Bowling Average Against a Team (as an All-Rounder)
bbi_allrounder: Best Bowling in Innings (as an All-Rounder)
bta: The number of balls to boundary against the player (Likely the number of balls bowled that resulted in boundaries)
bwa_allrounder:The number of balls with a wicket taken by the player (as an All-Rounder)

ct: Catches Taken (as a Fielder)
fw_allrounder: Five-Wicket Hauls (as an All-Rounder)
hn: Highest Score in an Innings (Context-Specific, could relate to batting)
hs: Highest Score in a Series
pr_allrounder: Player Rating (Specific to All-Rounder Performance)
rn: Runs (Context-Specific, likely related to runs conceded in bowling or scored in batting)
st: Stumpings Made (as a Wicketkeeper)
wk_allrounder: Wickets Taken (as an All-Rounder)
sp_allrounder: years active
name: Player Name
gender: Player Gender
bating_style: Batting Style
bowling_style: Bowling Style
playing_role: Playing Role (e.g., Batsman, Bowler, All-Rounder)
country_id: Country Identifier
country: Country Name

Appendix 2:
Final codes and datasets:
https://drive.google.com/drive/folders/1qUL0OCXgxB1B8MngCLn93tyBcQIs1snx?usp=drive_link