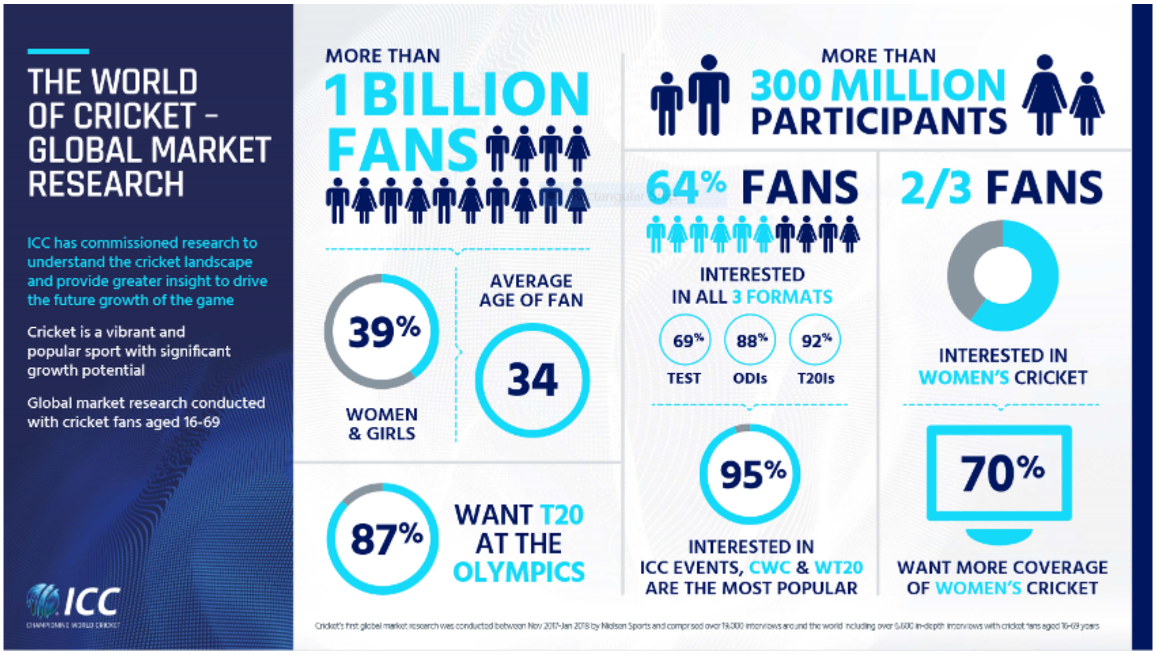
Beyond Scoreboards: Cricket Data Analysis and Interpretability of 'Player of the Match' Awards

SI 618 Final Project Report

Vidhi Bhatt (bvidhi), Prathamesh Joshi (prathuj)

Motivation


THE WORLD OF CRICKET – GLOBAL MARKET RESEARCH
ICC has commissioned research to understand the cricket landscape and provide greater insight to drive the future growth of the game
Cricket is a vibrant and popular sport with significant growth potential
Global market research conducted with cricket fans aged 16-69
ICC
MORE THAN 1 BILLION FANS
39% WOMEN & GIRLS
AVERAGE AGE OF FAN 34
87% WANT T20 AT THE OLYMPICS
MORE THAN 300 MILLION PARTICIPANTS
64% FANS INTERESTED IN ALL 3 FORMATS
69% TEST
88% ODIs
92% T20Is
95% INTERESTED IN ICC EVENTS, CWC & WT20 ARE THE MOST POPULAR
2/3 FANS INTERESTED IN WOMEN'S CRICKET
70% WANT MORE COVERAGE OF WOMEN'S CRICKET
Cricket's first global market research was conducted between Nov 2017-Jan 2018 by Nielsen Sports and comprised over 19,000 interviews around the world including over 6,600 in-depth interviews with cricket fans aged 16-69 years

Cricket, often hailed as a sport that blends skill, strategy, and precision, has undergone a transformative journey over the years. The motivation behind this data analysis project lies in the recognition of the pivotal role data plays in assessing cricketers and their skills. With the ongoing ICC Cricket World Cup 2023, everybody was in their sportsmanship and cheering loud and proud for their country, it felt rightfully so to work with the cricket data.

As the complexity of the game increases, teams and players are turning to data-driven insights to gain a competitive edge. This project seeks to delve into the statistical nuances of Test matches, One Day Internationals (ODIs), and Twenty20 (T20) matches, providing a comprehensive analysis that goes beyond the surface-level statistics.

By undertaking this project, we aim to contribute to the growing body of knowledge in sports analytics, specifically focusing on cricket. This endeavor is not isolated but draws inspiration from previous studies and reports that have successfully applied data analysis techniques to gain meaningful insights in the field of sports.

In conclusion, this project is driven by a compelling desire to unravel the intricacies of cricket through a data-centric lens. By offering a nuanced perspective on Test matches, ODIs, and T20s, we endeavor to empower teams, coaches, and enthusiasts with actionable insights that can shape the future of the game. As cricket embraces the era of data, this project aims to be a noteworthy contribution to the ever-expanding field of sports analytics.

Ref: https://www.icc-cricket.com/media-releases/759733

Literature Review:

Cricket data analysis has emerged as a thriving field, employing a diverse arsenal of statistical and machine learning techniques to decipher the intricacies of player performance, team strategies, and match outcomes. Analysts delve deep into player data like batting and bowling averages, strike rates, and dismissal types, unearthing strengths and weaknesses that might otherwise go unnoticed.

Cricket Data Analysis and Score Prediction (1) gives prediction of the match score based on strike rate, positions, average runs per match etc. Data Analytics in the Game of Cricket: A Novel Paradigm (2) endeavors to discover a novel paradigm of research in cricket analytics known as the timing index, based on a real life IoT-based implementation. Cricviz.com(3) is a website that provides cricket data and analytics to users, however we have not made use of any resources from this website.

End To End Cricket Data Analytics Project Using Web Scraping, Python, Pandas and Power BI (4) uses Power BI to perform analyses on T20 world cup cricket data.

Data Sources

Primary data -
The foundation of this comprehensive cricket data analysis project lies in the rich and extensive dataset procured from Kaggle. The dataset encompasses a wealth of statistical information on players' performance in One Day Internationals (ODIs), Test matches, and Twenty20 (T20) games, covering both batting and bowling categories.
Data: Link
Variables used:
The dataset comprises crucial metrics that illuminate the multifaceted aspects of a player's cricketing journey. These metrics include the player's career span, the number of matches played, innings batted or bowled, instances of not-outs, total runs scored, highest individual scores, batting average, balls faced, strike rate, centuries, half-centuries, and the total number of ducks.

Secondary data -
In tandem with the primary dataset from Kaggle, this project extends its analytical scope by incorporating supplementary data sourced from ESPN. ESPN, a leading sports media outlet, provides a wealth of information and statistics related to cricket, including player awards, player of the match accolades, and records across Test matches, One Day Internationals (ODIs), and Twenty20 Internationals (T20Is).
Data: Link
Variables used:
The supplementary dataset features key columns, including "Player Span" delineating the duration of a player's international career, "Mat" denoting the total number of matches played, and "Awards" highlighting instances where players have been honored with the prestigious Player of the Match recognition. Furthermore, the dataset segments the player's performances into Tests, ODIs, and T20Is, offering a nuanced breakdown of achievements across different formats.
The inclusion of awards as a metric adds a qualitative dimension to the analysis, allowing for the identification of standout individual performances that significantly impacted the outcome of matches. These accolades serve as markers of exceptional skill, contributing valuable context to the quantitative statistics derived from the primary dataset.

Self constructed Team acronym name table -
A manual compilation of team name acronyms was undertaken to facilitate accurate matching of team information.
Data: Link
I have kept this file in my personal Google Drive for anyone to view until December 31, 2023. The data is also printed in the Jupyter notebook for reference.
Variables used:

Acronym	Country
IND	India
SL	Sri Lanka

Data Manipulation

The success of any data analysis hinges on the quality and integrity of the underlying dataset. In this project, meticulous data cleaning processes were employed to refine the Kaggle dataset, ensuring accuracy and consistency across the board. The following steps outline the key measures taken to cleanse the data:

1. Handling Players with Identical Names:

- Seniority Marking: Players sharing the same name were identified and differentiated based on their seniority in the cricketing arena.
- Prioritization: For records with duplicate player names, a decision was made to retain the player with a superior record while eliminating redundant entries.

2. Cleaning Batting and Bowling Data:

- Name Segmentation: Player names were dissected into First Name, Last Name, and Team Name to enhance data granularity.
- Team Acronym Construction: A manual compilation of team name acronyms was undertaken to facilitate accurate matching of team information.

- Deduplication: Duplicate records within the files were systematically removed to streamline the dataset.
- Span Calculation: For players featuring in all three cricket formats, the player's career span was derived as the minimum of the start year and the maximum of the end year across formats.
- Numerical Summation: Numerical scoring columns in both batting and bowling datasets were aggregated by player information.

3. Merging Batting and Bowling Data:

Batting and bowling datasets were merged based on player information, consolidating the statistical insights across different facets of the game to be called ‘cricket’.

After aggregating the Batting and Bowling Data our Data got in sync with the Test matches | Individual records (captains, players, umpires) | Longest careers | ESPNcricinfo, for Longest careers of players which confirmed that the data aggregation was successful.

4. Merging Primary and Secondary DataSets

- The primary dataset serves as the features for Player-of-the-Match interpretation while the secondary dataset serves as labels for it. The players were combined across two datasets and we were able to retain 83 of the 85 records from the secondary dataset.
- Inorder to consider also the players with no awards, so as to balance the dataset for model prediction, we sampled 200 more samples from the remaining dataset.

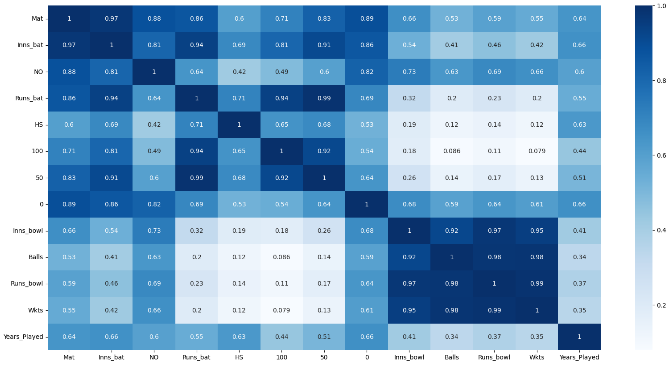
Visualisations

In the pursuit of unraveling the intricate dynamics of cricket through data analysis, a multifaceted approach was adopted. The following key analyses and visualizations were conducted, providing a comprehensive exploration of player and team performances:

We have given a Detailed interpretation of Observations and Insights for each visualization in the Jupyter Notebook supported with code.

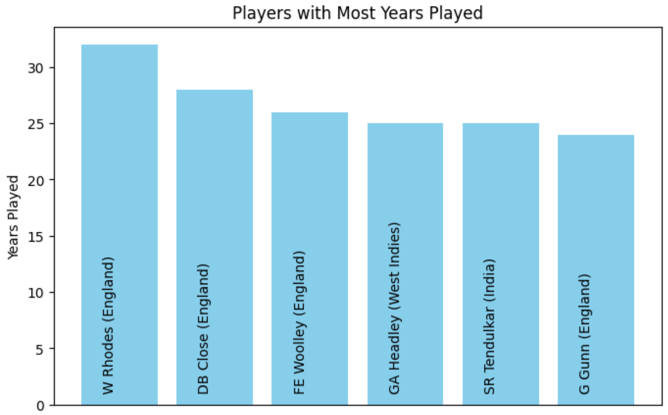
Correlation Matrix:

- A correlation matrix was constructed to unveil relationships between various cricketing metrics. This matrix serves as a valuable tool for identifying patterns and dependencies within the dataset, shedding light on the interplay between different statistical variables.

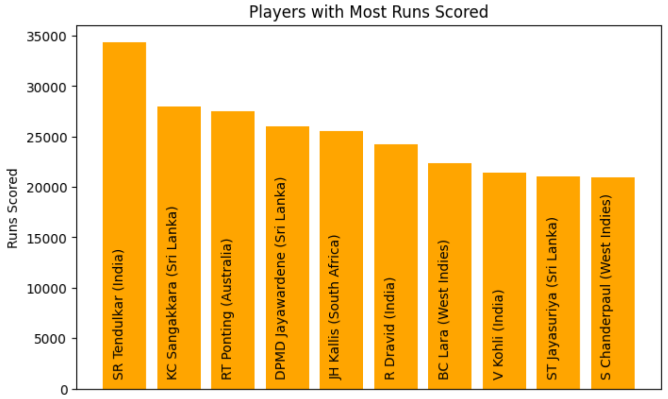


Player Analysis:

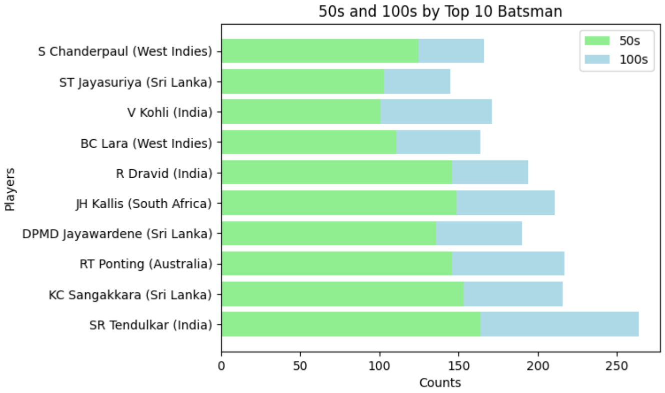
- Most Years Played: Identification of players with the longest careers, showcasing durability and longevity in international cricket.



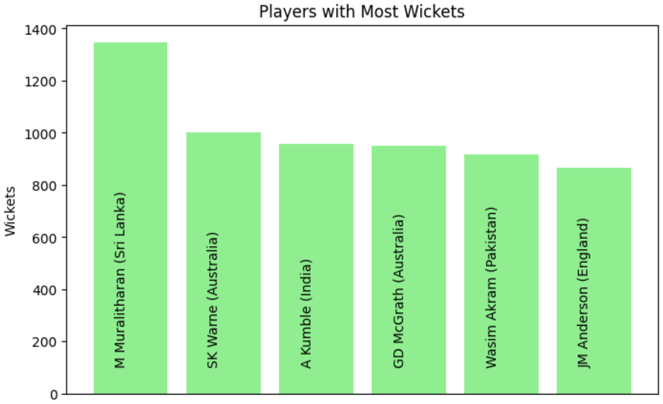
- Most Runs Scored: Recognition of players who have made substantial contributions with the bat, securing top positions in the run-scoring charts.



- Stacked Bar Chart of 50s and 100s: Visual representation of the number of half-centuries and centuries scored by the top 10 batsmen, providing insights into their consistency and ability to convert stats into big scores.

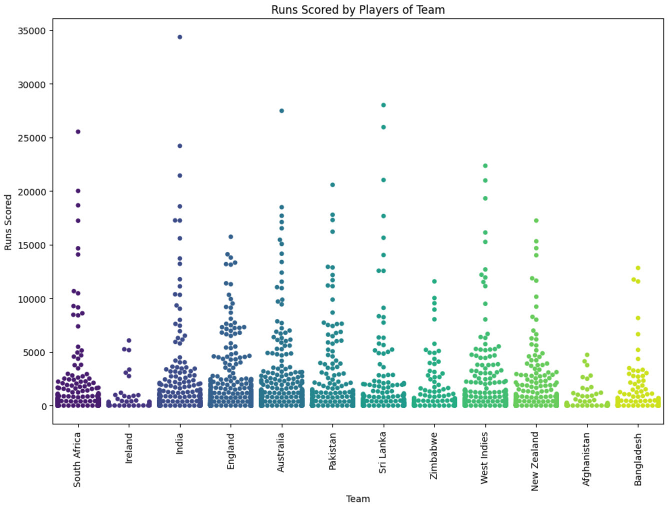


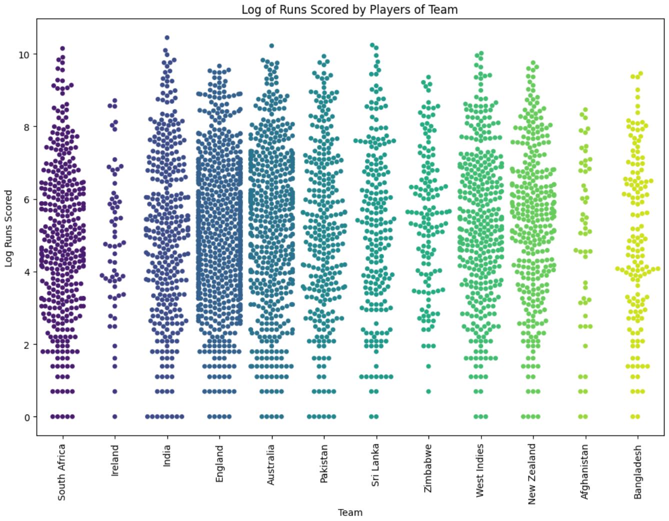
- Players with Most Wickets: Highlighting bowlers who have excelled in taking wickets, a crucial metric in assessing bowling prowess.



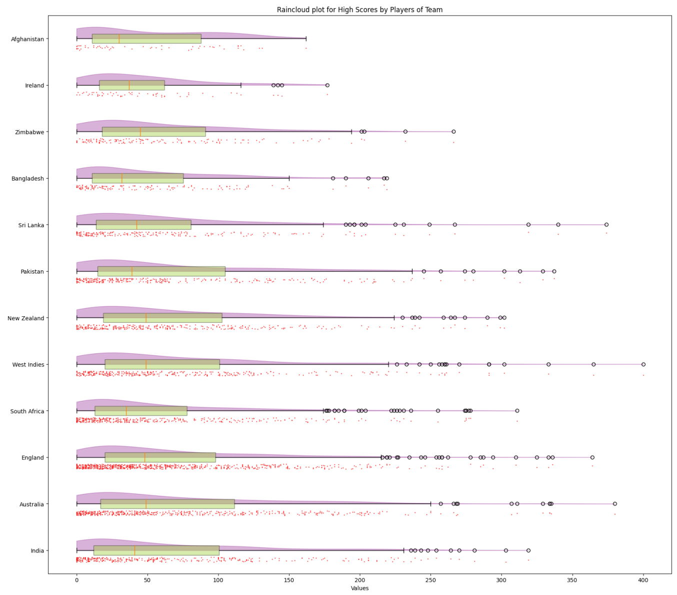
Team Analysis:

- Total Runs Scored Teamwise: Aggregation of runs scored by teams such as India, England, Australia, etc., offering a comparative perspective on batting performances across different cricketing nations.

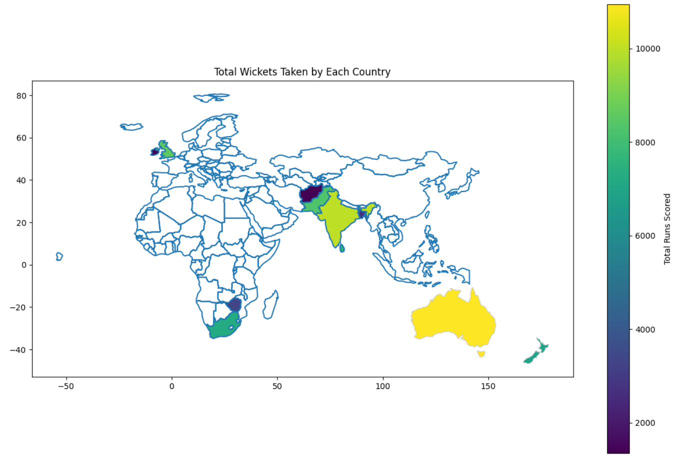




- Raincloud of Teams Using Runs: Utilization of a raincloud plot to showcase the distribution of runs scored by teams, allowing for a nuanced understanding of batting performances.

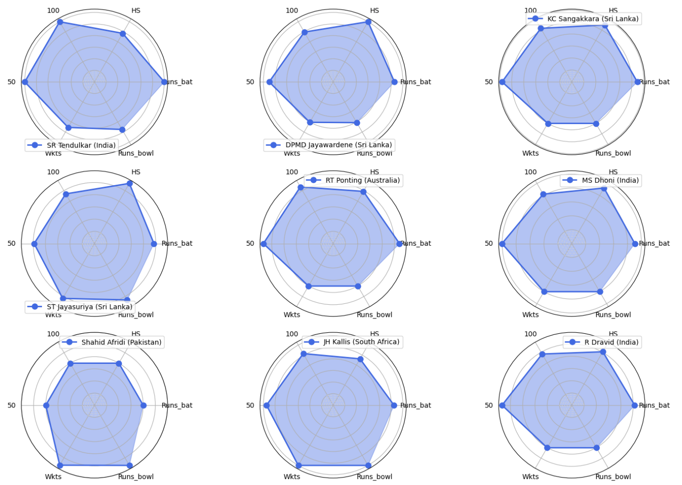


- World Plot of Teams with Most Wickets: A global visualization depicting teams that have collectively taken the most wickets. This plot provides a macroscopic view of bowling prowess across different cricketing nations.



Skill Plot

- Sub-Radar Plots for Players with Most Matches Played: Utilizing sub-radar plots to analyze specific skills such as 'Runs_bat', 'HS' (Highest Score), '100', '50', 'Wkts' (Wickets), and 'Runs_bowl' for players with the highest number of matches played. This granular examination allows for a detailed assessment of player proficiency in various facets of the game.

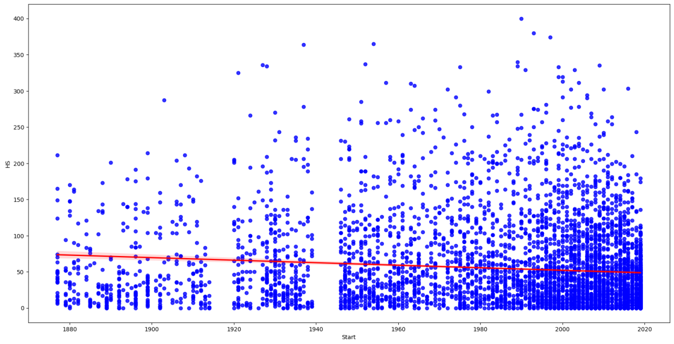


Data Analysis

Regression analysis

Unveiling the Relationship Between Highest Score and Career Start Year:

In an effort to discern any potential relationship between a player's highest score and the commencement of their cricketing career, a regression analysis using Ordinary Least Squares (OLS) was conducted. This statistical method allows us to explore whether a linear relationship exists between the two variables and ascertain the statistical significance of any observed trend.



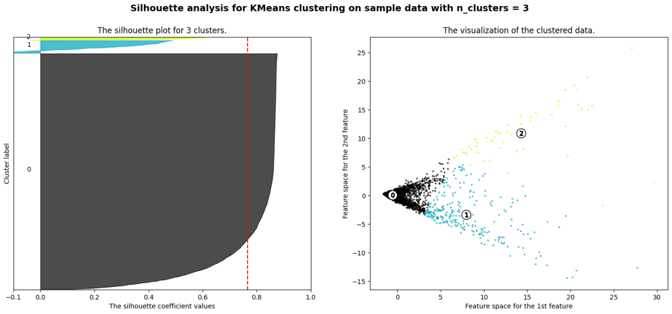
Although in some cases the Highest Score created by players has increased in time, this regression analysis was observed to be statistically significant and multiple reasons like match format changes, decreased years played and change in batting tendencies were inferred as the causative factors.

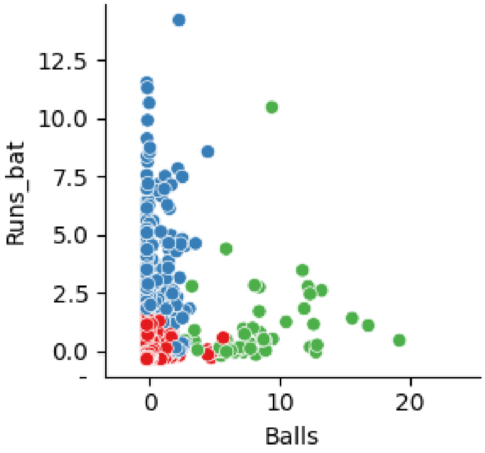
Clustering

As a general intuition to understand the structure of the cricket player's data based on their skills and abilities we make an attempt to categorize data into clusters and

attempt to understand if they reveal any hidden patterns to determining type of players. Accordingly, we cluster the players by KMeans clustering after transforming the data into 5 principal components and find that the best clustering Silhouette score is obtained for 3 clusters.
When we make a plot of the 3 clusters for different variables, we can see that for `Runs_bat` v/s `Balls` we can identify that the players are clustered as Good Batsman, Good Bowlers and intermediate players.

```
For n_clusters = 3 The average
silhouette_score is : 0.7658920034094178
For n_clusters = 4 The average
silhouette_score is : 0.6124250922087301
For n_clusters = 5 The average
silhouette_score is : 0.5591807953474736
```





Player of the Match Interpretation

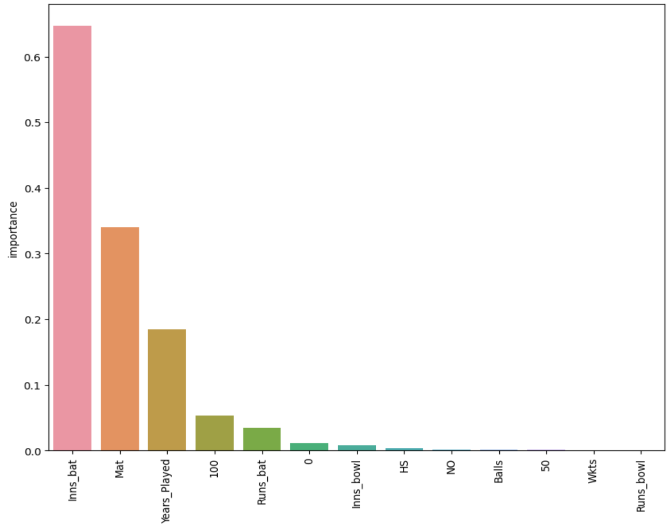
With the aim to understand the Interpretability of ‘Player of the Match’ Awards, we attempt to fit Machine learning models. Here our labels are the `Awards` from our Secondary data which is combined with the Primary data to obtain features. After we perform data cleaning and combining the two datasets we aim to fit regression models for our numerical outputs of ‘Awards’

Model Selection

We do a train-test split and fit Random forest regressor and Gradient Boosting Regressors to our data. For Random forest we obtain test accuracy of 89.6% and 91.7% test accuracy is obtained for Gradient Boosting Regressors. Given its superior performance compared to other regression models on this dataset and its ability to handle complex relationships and non-linearity in the data, Gradient Boosting Regressors is the natural choice of our model.

Feature Importance

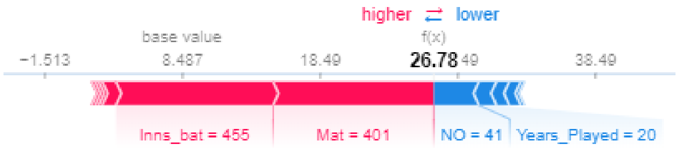
Permutation feature importance is a model-agnostic technique used to assess the relative importance of features in a machine learning model. It works by shuffling the values of a single feature and then measuring the change in the model's performance. Features that cause a larger drop in performance when their values are shuffled are considered more important.
For the purpose of our model fit and this data, ‘Inns_bat’ and ‘Mat’ are seen to be the most important features in determining the Player-of-the-Match awards.



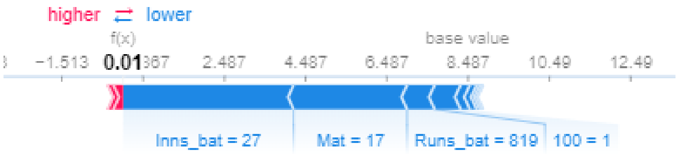
Player of the Match explanation by SHAP

SHAP (SHapley Additive exPlanations) values elucidate the influence of individual features on machine learning model predictions. They provide both local and global interpretability by quantifying the impact of each feature on specific predictions and across the dataset.

SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction we'd make if that feature took some baseline value.



For SR Waugh (Australia), SHAP explains how it predicted 26.78 as Player-of-the-Match score where the actual score was 26. It shows that while `Inns_bat` and `Mat` played a positive role in scoring the awards, `NO` and `Years_Played` contributed negatively towards it.



For AJ Pithey (South Africa), SHAP explains how it predicted 0 as Player-of-the-Match score and the actual score too was 0.

Conclusion

Overall, SHAP values offer a powerful lens into the inner workings of your player of the match prediction model. By providing granular insights into individual predictions, identifying key influencing factors, and even revealing potential biases, they can help you build a more interpretable, reliable, and ultimately, more trustworthy model.

Statement of Work

Vidhi Bhatt:

Contributed towards various data searches for project selection, primary data analysis and visualizations with expertise of utilization of various seaborn packages and report writing tasks.

Prathamesh Joshi:

Contributed towards data cleaning and manipulation tasks, data analysis for regression and clustering including applying models and working on Player of the Match explanation by SHAP.

Collaboration:

A very good collaboration was achieved through direct meetings, online Google Meets and shared Google Docs, thanks to University of Michigan's Information and Technology Services.

Disclaimer:

Utilized Github's Copilot resource for code using University of Michigan's Student Activation. The contributors are thankful to Professor Chris Teplovs and all the GSIs for their consistent efforts and feedback.

References

Research Papers:

1. Rahman, Sagidur. "Cricket Data Analysis and Score Prediction." Kaggle, 2018. https://www.kaggle.com/code/sazid28/cricket-data-analysis-and-score-prediction
2. Vishwarupe, Varad, et al. "Data Analytics in the Game of Cricket: A Novel Paradigm." Journal of Sports Sciences 41.1 (2023): 1-8. https://www.sciencedirect.com/science/article/pii/S1877050922008523

Websites:

3. CricViz. https://cricviz.com/
4. "End To End Cricket Data Analytics Project Using Web Scraping, Python, Pandas and Power BI." YouTube, uploaded by codebasics, 11 June 2022, https://m.youtube.com/watch?v=4QkYy1wANXA.