Matchup Mayhem: Data Mining for Fantasy Cricket Rivalry (IPL)

Het Vashi
Student, Indiana University

Laxmikant Kabra
Student, Indiana University

Vaibhav Lodhiya
Student, Indiana University

project-hetvashi-laxkabra-vlodhiya

Abstract

The Indian Premier League (IPL) is a Twenty20 cricket league in India, renowned for its high-paced, high-stakes matches. Fantasy cricket, a popular pastime among IPL and cricket enthusiasts, offers fans the opportunity to engage with cricket on a deeper level. The main challenge in fantasy cricket lies in selecting the best 11 from two teams that are playing a match such that the 11 aggregate maximum points which are given to each player that plays well on a given day. Our project presents a methodology to address this challenge by developing a technique for optimal player selection in IPL fantasy cricket. Our project involves the implementation of data-driven algorithms that consider player performance, pitch conditions, historical data, and various other factors to recommend the best 11 players for a given IPL match. These algorithms are designed to adapt to the ever-changing dynamics of IPL cricket matches, ensuring that users can consistently field high-performing fantasy teams.

Keywords

Logistic Regression, K-means clustering, fantasy cricket, Indian Premier League, Twenty20

1 Introduction

The Indian Premier League (IPL) stands as a pinnacle in the world of cricket, captivating audiences both domestically and internationally since its inauguration in 2008 under the auspices of the Board of Control for Cricket in India (BCCI). Operating on a franchise-based model, IPL teams represent diverse cities or regions, engaging in a round-robin group stage where each team competes twice against others in the tournament. Known for its high-intensity Twenty20 format, the IPL has not only garnered an immense fan following but has also left an indelible mark on cricket, blending sporting prowess with entertainment and glamour. The league's economic impact is substantial, encompassing revenue streams from broadcasting rights, sponsorships, advertising, and ticket sales.

In recent years, the burgeoning popularity of fantasy sports, particularly in cricket, has witnessed a surge in participant numbers. While fantasy cricket offers a compelling and enjoyable

experience for enthusiasts, the decision-making process often falters due to a lack of knowledge or personal biases. Consequently, there exists a significant hurdle in making informed and triumphant selections in fantasy cricket. This project endeavors to overcome this obstacle by harnessing the power of data mining to develop a robust recommendation system for crafting optimal fantasy cricket teams for matchups between any two teams.

Our primary objective is to recommend a set of players with the highest likelihood of delivering stellar performances. Leveraging datasets encompassing historical match statistics and ball-by-ball performances from the inaugural season in 2008 to the latest season in 2023, we introduce two innovative methods in this paper:

a) Performance-Based Scoring Metric Model: Utilizing a linear regression model as the foundational prediction framework, our approach involves devising a scoring metric model centered on individual player performances. This model considers various metrics such as consistency, balls faced, runs scored, wickets taken, strike rate, and more, tailoring the evaluation criteria to the unique roles of batsmen, bowlers, all-rounders, and keepers. The scoring model is honed by analyzing player data spanning the past five years.

b) In-Game Position-Based Optimal Cluster Prediction Using K-means Clustering: Our second method delves into the predictive power of k-means clustering, specifically tailored for in-game positions. With a dataset normalized and standardized to ensure uniformity in feature scales, we determine the optimal number of clusters, considering factors such as batsmen, bowlers, and all-rounders. Applying the k-means algorithm to preprocessed data, players are assigned to clusters based on the similarity of their attributes. Cluster interpretation involves analyzing defining features, such as batting average, bowling average, and other relevant statistics. The final player selection ensures a balanced team composition, considering the specific requirements for each playing category.

In summary, our project seeks to provide a comprehensive and data-driven approach to fantasy cricket team selection. Through meticulous analysis of player performances and innovative methodologies, we aim to empower fantasy cricket enthusiasts with the tools needed to make informed and successful choices for their dream teams.

Previous work

The referenced papers collectively tackle various challenges in the domain of cricket prediction, team selection, and performance evaluation using data mining and machine learning techniques. Balasundaram et al. [3] concentrate on developing a prediction model for classifying players and selecting cricket teams, especially in the ODI format. Hasanika et al. [7] utilize ball-by-ball data from sources like ESPN CricInfo to predict cricket match outcomes, employing decision tree algorithms and machine learning techniques. Kansal et al. [5] explore player valuations in the Indian Premier League (IPL) auction, analyzing determinants based on performance, experience, and other characteristics. Kumar et al. [6] predict outcomes of ODI matches using decision trees and MLP networks. Bonidia et al. [2] conduct a systematic review of data mining in sports, following established methodologies for comprehensive evaluations. Singh et al. [4] propose a model for predicting first innings scores and match outcomes in ODI cricket matches using Linear Regression and Naïve Bayes classifiers. Sumathi et al. [1] analyze cricket players' performance using linear regression, K-means, and random forest models, aiding in team formation. Thenmozhi et al. [8] focus on predicting IPL match outcomes using machine learning algorithms, achieving varying accuracies for different teams. Together, these papers contribute to the diverse landscape of cricket analytics, offering insights into team selection, player valuation, and match outcome prediction through a range of data-driven methodologies.

2 Methods

Our project evolved through various ideation phases, with ideas continuously shaping and refining. Following data collection from Kaggle and initial exploratory data analysis, we narrowed down our focus to two distinct model concepts. The selection of data was methodical, aligning with the specific requirements of each model and idea.

For the Scoring Model utilizing Linear Regression, we opted for a dataset that comprised ball-by-ball information for every match. This custom-curated dataset included features deemed crucial for quantifying a player's performance, aligning with the scoring criteria prevalent on cricket fantasy platforms. The chosen features were tailored to mirror the marking distribution on these fantasy platforms.

On the other hand, the Clustering Optimal Players using the K-means model necessitated a dataset containing career statistics for each player. We selected a dataset offering player statistics per season, ensuring it met the requirements for clustering players effectively. Both models underwent training utilizing data spanning the past five years. This deliberate choice aimed to mitigate the impact of outdated or irrelevant data, emphasizing recent performances to prevent bias in the models.

1. **Performance-Based Scoring Metric Model using Linear Regression:**

   The initial crucial step in this model involved crafting a meticulously curated custom dataset. With a focus on the ball-by-ball data, we carefully considered 33 columns, each representing data for every ball bowled in each match. This transformation was conducted separately for batsmen and bowlers, acknowledging the distinct parameters that define their performances.

   The batsmen dataset included player names, dates, match IDs, runs scored, balls faced, outs, and strike rates, while the bowling dataset encompassed match IDs, dates, bowler names, wickets, balls bowled, runs conceded, maidens, and economy rates. Both sets of data were instrumental in training our Linear Regression model.

   Significantly, the date played a pivotal role in our model. By extracting the year from the date, we engineered a feature called "time value." This time value assigned a weight to each data point, influencing the impact on the model. Older observations were weighted less compared to recent matches, ensuring minimal biases from a player's overall career. This rationale extended to our decision to focus only on the previous five years of match data, preventing the overall career performances from unduly influencing predictions for the current year.

   Recognizing the potential variation in the range of match statistics, we employed MinMaxScaler() for the Linear Regression models. Custom ranges were assigned to avoid introducing invalid values during subsequent calculations such as strike rate or economy.

   Post scaling, the next step involved creating the y-label for prediction, constituting our scoring metric. Random initial weights were assigned and manually fine-tuned to enhance model efficiency. Each weight, multiplied by the scaled value for each feature, contributed to a weighted score—our performance indicator for predicting a player's performance.

   For testing data, actual observed values from the 2023 test match season were unavailable. To address this, we calculated a five-year average for each player and used it as input for predicting player performance in upcoming matches.

   During training, we provided only basic features like runs, balls, wickets, etc., ensuring that the model learned the underlying relationships in the data without being influenced by our calculations and weights. Finally, the predicted values were sorted, player names were attached, and the results were arranged in descending order. The top-n players emerging

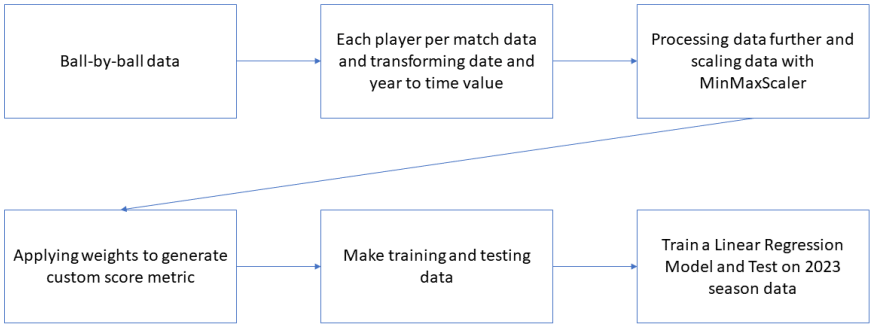


Figure 1: Model flow of linear regression model

from this process represented our best-predicted players for an upcoming match. figure 1 (a) shows a flow of complete model in six steps.

2. **In-Game Position-Based Optimal Cluster Prediction Using K-means Clustering:**

   We employed an unsupervised K-means algorithm to select the best 11 players for maximizing fantasy league points. In fantasy league scoring, points are assigned based on runs, the number of fours and sixes scored by batsmen, and for bowlers, it's the number of wickets, with points deducted for the runs they allow batsmen to score.

   In developing our clustering model, we aimed to avoid relying on a single feature, such as runs or wickets, to prevent inaccurate player selection. To address this, we introduced a weighted score, combining different features weighted according to their importance in achieving optimal player selection.

   The selected features were chosen carefully with points allocation in mind. For batsmen, crucial features included runs, the number of boundaries and sixes, and the batting average. While runs were readily available in our dataset, predicting the number of sixes and boundaries was challenging. However, this could be approximated by the batsmen's strike rate, indicating their scoring speed. Another essential feature was the number of innings played.

   For bowlers, key features were economy, wickets, the number of innings played, and the strike rate. Once the features were determined, the challenge was to assign appropriate weights for the weighted score, considering the different skills demanded by various player positions in a cricket team.

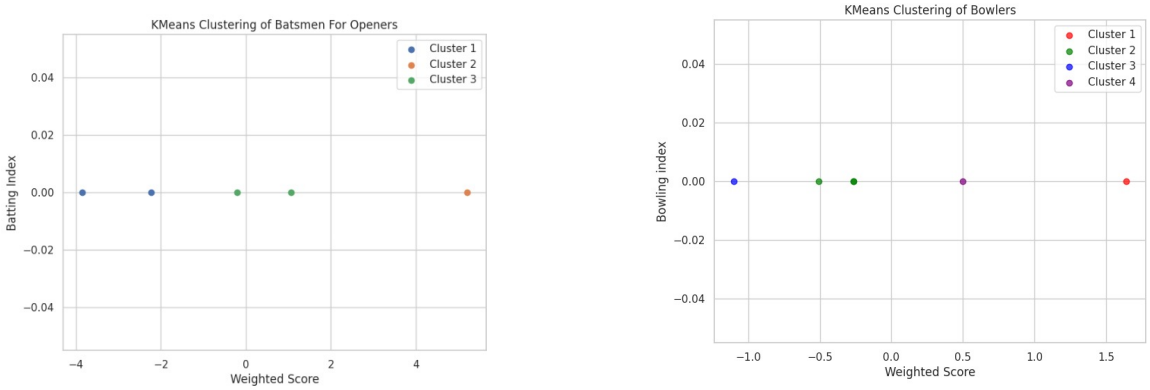
   Recognizing the distinct roles in a team—opening batsmen, middle order, lower order, and bowlers—we decided to cluster based on player positions. For instance, when selecting the best two opening batsmen, we clustered all opening batsmen from both teams. Since both strike rate and average were equally crucial for opening batsmen, we gave them equal importance in the weighted score. The same approach was applied to middle-order batsmen, with a slightly higher weight for average to reflect their stabilizing role in case the openers fail.

   Lower-order batsmen, responsible for scoring valuable runs with fewer balls left, had a higher emphasis on strike rate in the weighted score used for clustering. For bowlers,

wickets and strike rate held more significance than other features.

The determination of the number of clusters was a result of extensive experimentation. We settled on three clusters for each batting position and four clusters for bowlers. In the player selection process, the top cluster for batsmen was chosen based on the highest weighted score. If the top two batsmen couldn't be filled from the first cluster, we proceeded to the second cluster with the second-highest weighted score. For bowlers, the selection involved choosing the cluster with the lowest weighted score since, for bowlers, all parameters should be lower.

This methodology allowed for a nuanced and position-specific approach in player selection, contributing to a more accurate and strategic fantasy league team composition.



(a) Cluster for Batsmen

(b) Cluster for Bowlers

Figure 2: Clusters

3 Results(2023)

Matches	**K-means (Actual - Predicted)**	**Error (K-means)**	**Linear Regression (Actual - Predicted)**	**Error (Linear Regression)**
CSK VS GT	4	36%	3	27%
PBSK VS KKR	4	36%	5	45%
LSG VS DC	4	36%	4	36%

Table 1: Results Table

In the evaluation of our methodology, it is crucial to acknowledge the benchmark against which our predictions were measured. The actual performance of players, used for comparison, was derived from reputable online cricket platforms. These platforms are recognized for their accurate and real-time statistical data, providing a reliable standard for assessing the effectiveness of our fantasy cricket player selection approach.

The comparison between the actual performance and our model's predictions was conducted with meticulous attention to detail. The discrepancies and errors reported in our results table (Table 1) are based on the divergence between the players selected by our methodology and the actual best performers as determined by the online cricket platforms.

This alignment with external and widely accepted sources ensures the transparency and credibility of our evaluation process. It also underscores the practical applicability of our

methodology in the context of real-world fantasy cricket scenarios, as enthusiasts often refer to online platforms for performance assessments and player selections.

4 Discussion

Linear Regression:

The meticulous curation of a custom dataset derived from ball-by-ball data stands out as a key strength in our methodology. The careful selection of pertinent features for batsmen and bowlers individually underscores the model's ability to account for the distinct parameters defining their performances.

The incorporation of the "time value" feature, derived from match dates, represents a thoughtful addition to the model. This feature effectively tackles the challenge of mitigating biases from a player's overall career by assigning dynamic weights to historical data. Emphasizing recent performances within the last five years ensures the model's adaptability to the evolving nature of player careers.

The adoption of MinMaxScaler() for scaling values emerges as a pragmatic choice, preventing skewed influences from features with varying ranges. This approach ensures a balanced treatment of disparate features within the dataset.

Our strategy of calculating a five-year average for players when specific 2023 season data was unavailable showcases adaptability. This method enables predictions for upcoming matches based on a player's historical average, maintaining relevance in the absence of season-specific data.

The Linear Regression model yielded promising results, boasting an average accuracy of 80%. The success can be attributed to meticulous data generation methods, including careful scaling and transformation, which helped avoid errors and eliminate invalid values (null or inf). Notably, our approach introduced a novel method absent from existing literature reviews and Kaggle notebooks, establishing a clear path for utilizing data to meet our specific needs. The dataset comprised over 13,000 points, with the 5-year span for model training incorporating more than 5,000 data points.

The manual assignment of weights to generate the custom score introduces a potential area for improvement. Further exploration and testing could optimize these weights, enhancing the model's accuracy and robustness.

Additionally, there is potential for further experimentation with model training using a different number of years. While we opted for a 5-year span, extensive testing could reveal whether a shorter duration could lead to improved model performance. Rigorous testing in this regard would contribute to a deeper understanding of the model's dynamics and its sensitivity to varying training periods.

K-means Clustering:

Our clustering approach based on player positions (opening batsmen, middle order, lower order, and bowlers) proved effective in tailoring the selection criteria to the specific demands of each role. The use of weighted scores ensured that no single feature dominated the selection process, leading to a more comprehensive evaluation of player performance.

The visual representation of clusters for batsmen and bowlers (Figure 2) provides a clear insight into how players are grouped based on their performance metrics. This transparency aids fantasy cricket enthusiasts in understanding the rationale behind player selection and fosters trust in the methodology.

The results from the K-means clustering model showcased a promising approach to player selection, with errors ranging from 27% to 30%. While no model is perfect, these errors are acceptable in the context of the unpredictable nature of cricket matches.

Future Directions Refinement of Clustering

Future iterations of our methodology could involve fine-tuning the clustering algorithm to adapt to evolving player strategies and changing match dynamics. Continuous validation and adjustment of the weighted scores can further enhance the accuracy of player selection. Integration of Real-time Data

Incorporating real-time data, such as player form and current match conditions, can add a dynamic dimension to our model. This integration would enable more adaptive and responsive player recommendations. Ensemble Modeling

Exploring ensemble modeling techniques, combining the strengths of different algorithms, could potentially improve the overall robustness and accuracy of our fantasy cricket player selection system.

5 Author Contribution Statement

Our research project, "Matchup Mayhem: Data Mining for Fantasy Cricket Rivalry (IPL)," was a collaborative effort that drew upon the unique strengths and expertise of each team member. The contributions of each author are outlined below:

Clustering Algorithm Implementation: Het Vashi and Vaibhav Lodhiya took the lead in implementing the clustering algorithms, specifically the K-means algorithm for player Selection in batting and bowling respectively. His expertise in unsupervised learning greatly influenced the design and effectiveness of our player selection methodology.

Linear Regression Modeling: Laxmikant Kabra led the idea development and implementation of the Linear Regression model for batting as well as bowling models. With attention to detail in data scaling and transformation significantly contributed to the model's accuracy.

Collective Effort:

Data Processing and Feature Selection: All authors ensured the availability of clean and relevant data for subsequent analysis.

Methodology Design: All authors actively participated in the design and conceptualization of the methodology. The decision to cluster players based on positions, the selection of features, and the overall approach to player selection were outcomes of collaborative discussions.

Results Analysis: The analysis of results, including the interpretation of clustering outcomes and regression model performance, was a collective effort. Each author provided valuable insights and perspectives, contributing to the depth of our findings.

Paper Writing: All authors collaborated in drafting and refining the paper. From the abstract to the conclusion, each section benefited from the diverse skills and perspectives of the team.

The authorship of this paper reflects a harmonious collaboration where each member brought their unique skills to the table. The synergy of our efforts resulted in a robust methodology for fantasy cricket player selection, as presented in this research report.

References

[1] A Balasundaram, S Ashokkumar, D Jayashree, and S Magesh Kumar. Data mining based classification of players in game of cricket. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, pages 271–275, Sep. 2020.

[2] Dinithi Hasanika, Roshani Dilhara, Dulanjali Liyanage, Asitha Bandaranayake, and Sampath Deegalla. Data mining system for predicting a winning cricket team. In *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*, pages 92–97, Dec 2021.

[3] Prince Kansal, Pankaj Kumar, Himanshu Arya, and Aditya Methaila. Player valuation in indian premier league auction using data mining technique. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pages 197–203, Nov 2014.

[4] Jalaz Kumar, Rajeev Kumar, and Pushpender Kumar. Outcome prediction of odi cricket matches using decision trees and mlp networks. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pages 343–347, Dec 2018.

[5] Robson Parmezan Bonidia, Jacques Duilio Brancher, and Rosangela Marques Busto. Data mining in sports: A systematic review. *IEEE Latin America Transactions*, 16(1):232–239, Jan 2018.

[6] Tejinder Singh, Vishal Singla, and Parteek Bhatia. Score and winning prediction in cricket through data mining. In *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*, pages 60–66, Oct 2015.

[7] M Sumathi, S Prabu, and M Rajkamal. Cricket players performance prediction and evaluation using machine learning algorithms. In *2023 International Conference on Networking and Communications (ICNWC)*, pages 1–6, April 2023.

[8] D. Thenmozhi, P. Mirunalini, S. M. Jaisakthi, Srivatsan Vasudevan, V Veeramani Kannan, and S Sagubar Sadiq. Moneyball - data mining on cricket dataset. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pages 1–5, 2019.