Advanced Machine Learning Techniques for Optimizing Sports Team Composition: A Comprehensive Predictive Analytics Framework

Submitted for Business Analytics Research Project to Aston University.

Submitted in September 2024

By

Niranjan Gopalan,

Aston Business School, Aston University.

Master of Science in Business Analytics

Aston University
BIRMINGHAM UK

Declaration

I declare that I have personally prepared this research report titled "Comprehensive Analysis of Indian Premier League and Projecting the Optimal Squad for the 2025 Season." This work has not been submitted for any other degree or qualification, nor has it appeared in any previously published document. The research described here is my own, conducted personally unless otherwise stated. All sources of information are duly acknowledged through references. This study contributes original insights to cricket analytics, particularly in IPL team management and player selection strategies.

Acknowledgement

I would like to express my sincere gratitude to sports performance analysts, research engineers worldwide, whose research and insights have significantly informed and enriched my understanding of sports analytics. I am deeply thankful to my mother Latha, for her unwavering encouragement and belief in my abilities, and to my father Gopalan, whose dedication to public welfare and mathematics literacy has been a constant source of inspiration. I also extend my heartfelt appreciation to my supervisor Dr. Rizwan Ahmed, for his guidance and support throughout this research project. Lastly, I would like to thank all the individuals who supported me during my research, as their contributions have been essential to the completion of this work.

Aston University
BIRMINGHAM UK

Table of Contents

Aston University
BIRMINGHAM UK

Abstract

This research project focuses on optimizing squad compositions for the Kolkata Knight Riders (KKR) and Delhi Capitals (DC) in preparation for the 2025 Indian Premier League (IPL) mega auction. The study employs advanced cricket analytics and machine learning models to provide strategic insights for team building and performance enhancement. Since its inception in 2008, the IPL has revolutionized cricket, becoming one of the most popular and lucrative sports leagues globally, featuring ten franchise teams competing in a high-stakes, fast-paced T20 format. The research methodology involves comprehensive data processing, including extraction from reliable sources, cleaning, and preprocessing. Various machine learning models, such as linear regression, random forest, XG boosting, and support vector regression, are utilized to analyse player performance and predict outcomes. Key analyses include Win Ratio Analysis of Teams, a rule-based scoring system combining normalisation and weighted aggregation, and Random Forest Models optimized using RandomizedSearchCV. The study evaluates these models using performance metrics like ROC curves and performance distribution curves to ensure robust and accurate predictions. By analysing KKR's championship-winning strategies in 2024 and DC's approach to team building, the research provides a comparative analysis of different management philosophies and their impact on team performance.

The study's significance extends beyond the IPL, offering valuable insights for other T20 leagues like The Hundred and Big Bash League (BBL), as well as potential applications in sports like football and baseball. Key findings highlight the importance of strategic player retention, the influence of external factors on performance, and the challenges of predicting player auction values. For KKR and DC specifically, the research offers analysis of current player performances, identification of key strengths and weaknesses, and recommendations for optimizing squad potential, with a particular focus on helping DC better utilize their young talent. The research contributes to the growing field of quantitative sports analytics, demonstrating the importance of data-driven decision-making in modern sports management. It provides a framework for improving player selection, strategy formulation, and overall team management across various sports disciplines. Limitations of the study include the challenges of evaluating new players with limited IPL data, accurately predicting auction values, and accounting for unforeseen circumstances. In conclusion, this research project offers a comprehensive analysis of cricket analytics, emphasizing the importance of data-driven strategies in sports management. By focusing on the squad optimization of KKR and DC, the study provides valuable insights that can be applied to other cricket tournaments and sports, underscoring the potential of analytics to revolutionize team management and performance optimization in the competitive world of sports.

Keywords: Indian Premier League, Kolkata Knight Riders, Delhi Capitals, Squad Optimization, Player Performance, Machine Learning, Data Analysis, Predictive Modelling, Win Ratio, XG Boosting, Random Forest, Batting Statistics, Bowling Statistics, Performance Metrics, Auction Strategies, Team Management, Cricket Analytics, Statistical Methods, Player Selection, Team Performance

Word count: Around 11,200 words

Aston University
BIRMINGHAM UK

List of Figures

Aston University
BIRMINGHAM UK

List of Tables

1. Introduction

1.1 Background

The Indian Premier League (IPL) has revolutionized cricket since its inception in 2008, becoming one of the most popular and lucrative sports leagues globally (Board of Control for Cricket in India, 2023). This professional Twenty20 cricket tournament features ten franchise teams representing different Indian cities or states, competing in a high-stakes, fast-paced format that has captured the imagination of fans worldwide.

Figure 1- IPL Logo

Source: iplt20

The IPL's success can be attributed to several factors. Firstly, its star-studded lineups attract top cricket talent from around the world. Each team can field up to four overseas players in their playing eleven, creating a melting pot of international stars alongside India's best cricketers (ESPN Cricinfo, 2023). This combination of global and local talent has helped the IPL become a cricketing spectacle that consistently ranks among the top sports leagues in terms of average attendance.

The tournament's economic impact has been substantial. In 2022, the league's brand value was estimated at ₹90,038 crore (US$11 billion) (Duff & Phelps, 2022). Its contribution to India's GDP is significant, with the 2015 season alone adding ₹1,150 crore (US$140 million) to the economy (BCCI, 2016). The league's valuation has skyrocketed, reaching US$10.9 billion in December 2022 and achieving "decacorn" status (Economic Times, 2023).

The IPL's popularity is reflected in its lucrative media rights deals. For the 2023-2026 seasons, the league sold its media rights for US$6.4 billion, valuing each match at $13.4 million (Sportstar, 2023). The tournament has also broken viewership records, with the 2023 final becoming the most streamed live event on the internet, attracting 32 million viewers (JioCinema, 2023).

The Indian Premier League has transformed cricket from a traditional sport into a global entertainment spectacle. Its blend of star power, economic impact, and innovative gameplay has cemented its position as a powerhouse in the world of sports, influencing the way cricket is played and consumed around the globe (“see Appendix 1”).

Each team can have a maximum of 25 players in their squad, with no more than eight overseas players. The playing eleven for each match can include up to four overseas players, ensuring a balance of international stars and domestic talent (IPL Governing Council, 2024).

The IPL's team structure, with its mix of international stars and Indian talent, creates a unique and exciting cricketing spectacle that has captured the imagination of fans worldwide (Shah, 2023).

1.2 Team performances over the years

Table 1 - Team performance table

Team Name	Played	Won	Lost	N/R	Titles	Finalists	Playoff
MI	261	144	117	0	5	6	11
RCB	256	123	129	4	0	3	9
KKR	252	131	120	1	3	4	7
DC	252	115	135	2	0	1	6
PK	246	112	134	0	0	1	2
CSK	239	138	99	2	5	10	13
RR	222	112	107	3	1	2	5
SRH	182	88	94	0	1	3	6
GT	45	28	17	0	1	2	2
LSG	44	24	19	1	0	0	2

This table provides a comprehensive overview of the performance of Indian Premier League (IPL) teams since the league's inception in 2008.The breakdown of the information and analyse the data for each team is explained in the “Appendix 2”.

This table highlights the varying degrees of success and consistency among IPL teams. While some teams like MI and CSK have dominated with multiple titles and consistent playoff appearances, others like RCB and PK have struggled to convert their opportunities into championships. The newer teams, GT and LSG, have shown promise in their short IPL careers, adding excitement to the league's competitive landscape (Shah, 2023).

1.3 Research Objective and Need for Study:

A major focus of this study is the upcoming mega auction for the 2025 IPL season. This auction will result in most players being released, with teams allowed to retain only four players, including a maximum of two foreign players. This significant event provides an opportunity to analyse and optimize squad-building strategies for **Kolkata Knight Riders (KKR) and Delhi Capitals (DC)**.

Among the 10 teams in the Indian Premier League (IPL), this study aims to analyse, predict, and optimize the squads for two specific teams: Kolkata Knight Riders (KKR) and Delhi Capitals (DC). KKR, the 2024 IPL champions, boasts one of the strongest squads in the league. In contrast, DC possesses a power-packed young squad but has struggled to utilize their potential effectively. The research will focus on the following aspects:

1. Squad Analysis: Examine the composition of both KKR and DC squads, identifying key strengths and weaknesses based on their performances.
2. Quantitative analysis: Use statistical methods to compare player performances, team strategies, and match outcomes for both KKR and DC.
3. Performance Prediction: Develop models to forecast player and team performance based on historical data and current squad dynamics.
4. Squad Optimization: Propose strategies for both teams to maximize their squad potential, with a particular emphasis on helping DC better utilize their young talent.

5. Evaluate the current squads of KKR and DC to identify potential retention candidates.
6. Analyse the impact of retaining only four players on team dynamics and performance.
7. Success Factors: Investigate the elements that contributed to KKR's championship win in 2024, including team balance, leadership, and player utilization.

By conducting this research, the aim is to provide valuable insights into effective squad building, talent utilization, and performance optimization in the highly competitive environment of the IPL. The findings could offer strategic guidance not only for KKR and DC but also for other T20 cricket franchises globally.

1.4 Scopes

1. Player performance prediction: Develop models to predict player performance based on historical data, considering factors like batting and bowling statistics.
2. Team efficiency analysis: Evaluate the efficiency of teams using techniques like Data Envelopment Analysis (DEA) and Structural Equation Modeling (SEM).
3. Strategic player retention: Analyse strategies for the upcoming mega auction, focusing on optimal player retention decisions for KKR and DC.
4. Impact of external factors: Examine how factors like weather, match location, and stadium conditions affect player and team performance.
5. Comparative analysis: Compare KKR and DC's squad building and utilization strategies with other successful IPL teams.

1.5 Limitations of the study:

1. Limited data for new players: Relying solely on IPL data may limit the evaluation of new or emerging players who haven't played in the league before.
2. Complexity of player valuation: Accurately predicting player auction values and performance can be challenging due to multiple influencing factors.
3. Changing league dynamics: The study's findings may be affected by evolving league rules, team strategies, and player availability.
4. External factors: Unforeseen circumstances like injuries, player form, or off-field issues can impact team performance and are difficult to account for in models.
5. Limited scope: Focusing on only two teams (KKR and DC) may limit the generalizability of findings to other IPL teams or T20 leagues.
6. Time constraints: The dynamic nature of T20 cricket and frequent player transfers may make long-term predictions challenging.
7. Multi-objective optimization: It is difficult to formulate team selection as a multi-objective optimization problem, while considering budget constraints.

These scopes and limitations can help, frame the research objectives and methodology for analysing and optimizing the squads of KKR and Delhi Capitals in the context of the upcoming IPL mega auction.

2. Significance of the Study

1. Performance Optimization: By analysing the factors contributing to KKR's success and DC's underperformance, the study can offer valuable insights into how teams can better utilize their player resources, especially young talent (Ishi et al., 2022).

2. Quantitative Sports Analytics: The research contributes to the growing field of quantitative sports analytics in cricket, which has become increasingly important for team management and strategy development (Jana et al., 2021).

3. Player Improvement: The insights gained from this study can help improve individual player performance by identifying key areas for development based on data-driven analysis. (Techiexpert, 2024).

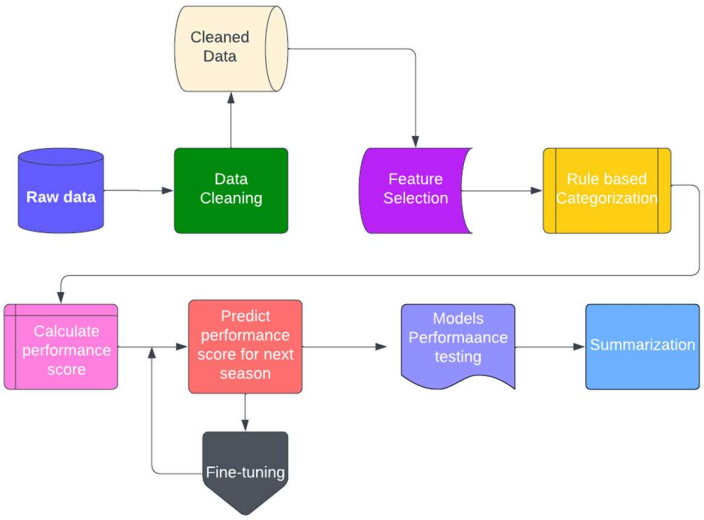
4. Global Cricket Applications: The findings could be valuable not only for IPL teams but also for national cricket boards such as the England and Wales Cricket Board (ECB), Cricket Australia, and New Zealand Cricket, helping them in player selection and team strategy for international competitions (Kalgotra et al., 2014).

5. Predictive Modelling: This research will contribute to the development of more accurate predictive models for player and team performance in T20 cricket, drawing inspiration from advanced analytics techniques used in football. These models can leverage machine learning algorithms and big data analysis, like those employed in predicting football match outcomes and player performance (Hubáček et al., 2019; Berrar et al., 2019). Such approaches can be valuable for team management, fantasy cricket enthusiasts, and sports analytics professionals, potentially improving decision-making processes in player selection and strategy formulation.

6. Comparative Analysis: By conducting an in-depth comparison of Kolkata Knight Riders' championship-winning strategies in 2024 and Delhi Capitals' approach to team building, this study will provide valuable insights into the efficacy of different management philosophies and their impact on team performance in the IPL (ESPNcricinfo, 2024). This analysis will highlight how KKR's meticulously designed squad, enabling aggressive batting without compromising depth, contrasts with DC's focus on nurturing young talent, offering a comprehensive perspective on successful team construction in T20 cricket.

2.1 Structure of the research

The goal of this study is to predict the optimal squad composition for KKR and DC in the upcoming Indian Premier League (IPL) season. The research methodology encompasses several key steps:



1. Data Extraction: Gathering comprehensive player statistics and performance data from various reliable sources.
2. Data Cleaning: Preprocessing the collected data to ensure accuracy, consistency, and relevance for analysis.
3. Descriptive Analytics: Conducting a thorough exploratory data analysis to understand the underlying patterns and trends in player performances.
4. Data Visualisation: Creating insightful visual representations of the data to facilitate easier interpretation and identification of key insights.
5. Feature Engineering: Developing new variables or transforming existing ones to enhance the predictive power of the models.
6. Player Score Prediction: Utilizing advanced statistical and machine learning techniques to forecast individual player performances based on historical data and relevant factors.
7. Hyperparameter Tuning: Optimizing the predictive models through rigorous hyperparameter adjustment to improve accuracy and reliability.
8. Model Performance Testing: Evaluating the performance of the tuned models using metrics and techniques such as cross-validation, ROC curves, and performance distribution curves to ensure robustness and accuracy.
9. Squad Optimization: Employing the refined predictive models to determine the most effective squad compositions for KKR and DC, considering various constraints.
10. Stakeholder Recommendations: Formulating data-driven, actionable recommendations for team management, coaches, and other relevant stakeholders to inform their decision-making processes in player selection and team strategy.

This comprehensive approach aims to leverage advanced analytics to provide valuable insights and strategic advantages in the highly competitive landscape of the IPL.

3. Literature Review

The Indian Premier League (IPL) has not only revolutionized cricket but has also become a fertile ground for sports analytics since its inception in 2008. As the league has grown in stature, so has the sophistication of the analytical approaches used to understand and predict its dynamics.

The Economic Catalyst

(Kadapa,2013) highlighted the IPL's massive economic footprint, underscoring the financial imperative driving the adoption of advanced analytics. With billions at stake, teams and stakeholders are increasingly turning to data-driven approaches to gain a competitive edge.

Evolution of Analytical Approaches

The journey of IPL analytics has been one of continuous refinement. (Shah et al, 2016) laid important groundwork with their comprehensive analysis of IPL data from 2008 to 2015. Their work demonstrated the potential of machine learning in decoding the complexities of T20 cricket, setting the stage for more advanced studies. Building on this foundation, Prakash et al. (2019) developed a nuanced player ranking system using machine learning algorithms. Their model's success in predicting player rankings with high accuracy showcased the power of analytics in informing team selection strategies.

The Human Element in Data

While numbers are at the heart of analytics, recent research has emphasized the importance of translating data into actionable insights. (Ishi et al,2022) took a significant step in this direction by using machine learning for player classification. Their work helps bridge the gap between raw data and on-field strategy, providing coaches and managers with a more intuitive understanding of player capabilities.

Predictive Power and Its Limitations

The holy grail of sports analytics is accurate prediction, and IPL research has made significant strides in this area. (Amala Kaviya et al,2020) achieved an impressive 81% accuracy in predicting match outcomes. However, as any cricket fan knows, the game's unpredictability is part of its charm. These models, while powerful, serve as tools to inform decision-making rather than crystal balls.

Transparency in Analytics

Recognizing the need for interpretable results, (Bajaj,2023) explored the use of Explainable AI techniques. This approach not only predicts performance but also elucidates the factors influencing these predictions, making the insights more accessible and actionable for non-technical stakeholders.

Visualising Success

In the fast-paced world of T20 cricket, the ability to quickly grasp complex information is crucial. (Rodrigues et al,2019) addressed this need by focusing on data visualisation techniques. Their work highlights how effective visual representation can transform raw data into strategic insights, accessible to everyone from analysts to players.

The Road Ahead

1. Real-time analytics during matches could revolutionize in-game decision-making.

2. Integration of non-traditional data sources, such as social media sentiment and player biometrics, may provide a more holistic view of performance.
3. More sophisticated player valuation models could transform auction strategies.
4. The application of deep learning to video analysis promises to unlock new insights into player techniques and strategies.

“The field of IPL analytics is not just about numbers; it's about enhancing the beautiful game of cricket”. As analytics continue to evolve, they promise to enrich our understanding and enjoyment of the sport, providing fans, players, and managers alike with new perspectives on the game we love.

3.1 Adoption of Machine Learning on Sports

The adoption of Machine Learning (ML) in sports has seen significant growth in recent years, revolutionizing various aspects of athletic performance, strategy, and management.

Performance Analysis and Prediction:

ML has been extensively applied to analyse and predict athletic performance. (Ofoghi et al, 2013) demonstrated the use of ML algorithms to predict medal-winning performances in sprint kayaking, achieving an accuracy of 80%. Similarly, (Bunker and Thabtah, 2019) reviewed ML applications in predicting outcomes of various sports, finding that ensemble methods often outperform individual algorithms in accuracy.

Injury Prediction and Prevention:

A critical area where ML has shown promise is in injury prediction and prevention. (Rossi et al, 2018) developed a ML model to predict injuries in soccer players, achieving an accuracy of 80% in identifying high-risk athletes. Building on this, (Rommers et al, 2020) used ML techniques to predict injuries in youth soccer players, demonstrating the potential of these methods in protecting young athletes.

Tactical Analysis:

ML has transformed tactical analysis in team sports. (Memmert and Raabe, 2018) explored how ML algorithms can analyse complex patterns in soccer matches, providing coaches with insights that were previously unattainable through traditional methods. In basketball, (Cervone et al,2016) used ML to evaluate decision-making in real-time, offering a new perspective on player effectiveness beyond traditional statistics.

Player Recruitment and Scouting:

The application of ML in talent identification and recruitment has gained traction. (McHale et al,2012) developed a ML model to assess player performance in soccer, which has implications for scouting and transfer decisions. More recently, (Liu et al, 2020) used deep learning techniques to analyse player movements in basketball, providing a data-driven approach to talent evaluation.

Fan Engagement and Business Operations:

ML has also found applications in enhancing fan engagement and optimizing business operations in sports. (Fried and Mumcu, 2016) explored how ML can be used to personalize fan experiences and improve marketing strategies in professional sports. In ticket pricing, (Kemper and Breuer, 2016)

demonstrated how ML algorithms can optimize dynamic pricing strategies, potentially increasing revenue for sports organizations.

Challenges and Ethical Considerations:

Despite its potential, the adoption of ML in sports faces several challenges. (Caya and Bourdon, 2016) highlighted issues of data quality and interpretation in sports analytics, emphasizing the need for domain expertise in developing ML models. Ethical considerations have also come to the forefront, with (Loland, 2018) discussing the implications of ML on fairness and integrity in sports.

Future Directions:

The future of ML in sports looks promising, with several emerging areas of research. Wearable technology and IoT devices are expected to provide more granular data for ML models, as explored by (Seshadri et al., 2019) in their work on real-time performance tracking. Additionally, the integration of computer vision with ML, as demonstrated by (Thomas et al, 2017) in their analysis of tennis player movements, opens new avenues for automated performance analysis.

3.2 Adoption of Machine Learning on Cricket

The adoption of Machine Learning (ML) in cricket analytics has gained significant traction in recent years, with researchers from Europe and the USA contributing to this field. Here's a literature review focusing on key aspects:

1. **Match Outcome Prediction:**
   Researchers have applied ML techniques to predict cricket match outcomes. A study from the UK focused on English County twenty-over cricket matches, investigating the degree to which it's possible to predict match outcomes using ML algorithms. This research demonstrates the growing interest in applying advanced analytics to cricket.

2. **Performance Analysis:**
   ML has been used to analyse player and team performance in cricket. While not specifically focused on cricket, McHale et al. (2012) developed ML models to assess player performance in soccer, which has implications for similar applications in cricket, particularly for scouting and team selection strategies.

3. **Data-Driven Decision Making:**
   The adoption of ML in cricket analytics aligns with broader trends in sports analytics. Beal et al. (2019) conducted a comprehensive survey on artificial intelligence for team sports, which included cricket. They noted that ML methods have been applied to various aspects of sports, including tactical analysis and performance prediction.

4. **Challenges and Limitations:**
   While ML shows promise in cricket analytics, researchers have noted challenges such as the need for high-quality data and the complexity of cricket's rules and playing conditions, which can affect model accuracy (Beal et al., 2019). The dynamic nature of cricket, with its multiple formats and varying conditions, presents unique challenges for ML applications.

5. **Future Directions:**
   Ongoing research is focusing on improving the accuracy of predictive models and expanding the range of applications for ML in cricket. This includes real-time analysis during matches and more sophisticated player valuation models (Beal et al., 2019). The potential for ML to

enhance decision-making in areas such as team selection, strategy formulation, and player development is significant.

6. **Interdisciplinary Approach:**
   The literature suggests that successful adoption of ML in cricket requires an interdisciplinary approach, combining expertise in data science, sports science, and domain-specific knowledge of cricket (Beal et al., 2019).

The adoption of ML in cricket analytics is growing, the field is still evolving. Researchers continue to refine methodologies and explore new applications to enhance the understanding and analysis of the sport. The potential for ML to transform various aspects of cricket, from player performance analysis to strategic decision-making, is significant, but challenges remain in terms of data quality, model interpretability, and practical implementation.

3.3 Summary of Literature Review:

1. **Match Outcome Prediction:**
   Researchers have applied ML techniques to predict cricket match outcomes. A study focused on English County twenty-over cricket matches investigated the degree to which it's possible to predict match outcomes using ML algorithms. This demonstrates the growing interest in applying advanced analytics to cricket.

2. **Performance Analysis:**
   ML has been used to analyse player and team performance in cricket. While not specifically focused on cricket, studies like McHale et al. (2012) developed ML models to assess player performance in soccer, which has implications for similar applications in cricket, particularly for scouting and team selection strategies.

3. **Data-Driven Decision Making:**
   The adoption of ML in cricket analytics aligns with broader trends in sports analytics. Beal et al. (2019) conducted a comprehensive survey on artificial intelligence for team sports, which included cricket. They noted that ML methods have been applied to various aspects of sports, including tactical analysis and performance prediction.

4. **Challenges and Limitations:**
   Researchers have noted challenges such as the need for high-quality data and the complexity of cricket's rules and playing conditions, which can affect model accuracy (Beal et al., 2019). The dynamic nature of cricket, with its multiple formats and varying conditions, presents unique challenges for ML applications.

5. **Future Directions:**
   Ongoing research is focusing on improving the accuracy of predictive models and expanding the range of applications for ML in cricket. This includes real-time analysis during matches and more sophisticated player valuation models.

6. **Interdisciplinary Approach:**
   Successful adoption of ML in cricket requires an interdisciplinary approach, combining expertise in data science, sports science, and domain-specific knowledge of cricket (Beal et al., 2019).

7. **Emerging Technologies:**
   The European Cricket Network has partnered with Full track AI, an advanced machine

learning and artificial intelligence service, to provide ball tracking graphics, pitch maps, speeds, and other key data points using mobile phone technology (Emerging Cricket, 2023).

In conclusion, while the adoption of ML in cricket analytics is growing, the field is still evolving. Researchers continue to refine methodologies and explore new applications to enhance the understanding and analysis of the sport. The potential for ML to transform various aspects of cricket, from player performance analysis to strategic decision-making, is significant, but challenges remain in terms of data quality, model interpretability, and practical implementation.

4. Research Methodology

The primary objective of this research is to leverage machine learning techniques to predict and optimize the squad compositions for two Indian Premier League (IPL) teams: Kolkata Knight Riders (KKR) and Delhi Capitals (DC). This study aims to utilize a comprehensive approach that incorporates various statistical methods, algorithms, and predictive models to analyse player performance data. By employing advanced data mining techniques, feature engineering, and machine learning algorithms such as decision trees, random forests, and support vector machines (Regression), the research seeks to identify the most effective player combinations for each team. The goal is to provide data-driven insights that can inform team management decisions, particularly in the context of player selection for upcoming seasons and auctions.

4.1 Dataset and Approach Overview:

The dataset is taken from trusted websites: Cricsheet and Howstat. The dataset appears genuine, and cross-verification has been performed to check the legitimacy of the data. The dataset contains a ball-by-ball record of IPL matches, providing detailed information about each delivery, including match details, player information, runs scored, extras, and dismissals.

Table 2 - Dataset Columns

Column Name	Description
match_id	Unique identifier for each match
season	The IPL season year
start_date	The date the match started
venue	The location where the match was played
innings	The innings number (1st or 2nd)
ball	The ball number within the over
batting_team	The team currently batting
bowling_team	The team currently bowling
striker	The batsman facing the current ball
non_striker	The batsman at the other end
extras	Total extra runs scored on this ball
wides	Number of wide balls
noballs	Number of no balls
byes	Number of byes
legbyes	Number of leg byes
penalty	Any penalty runs awarded
wicket_type	Type of dismissal if a wicket fell
player_dismissed	Name of the player dismissed (if applicable)
other_wicket_type	Secondary wicket type (if applicable)

This table represents a comprehensive dataset used for analysing cricket matches, specifically focusing on the Indian Premier League (IPL). Each row in the table corresponds to a specific delivery (ball) in a match, providing detailed information about the events occurring during that delivery("see Appendix 5")

4.2 Data Processing

4.2.1 Data Filtering

1. The data is filtered to include only innings 1 and 2, excluding super overs.
2. Further filtering is applied to select only the seasons from 2021 to 2024.

4.2.2 Player Data Extraction

1. Unique player names are extracted from the 'striker', 'non_striker', and 'bowler' columns.
2. A Data Frame containing all unique player names is created.

4.3 Batting Statistics Calculation

1. Pivot tables are created to calculate runs scored and balls faced by each player in each season.
2. The pivot tables are merged to create a comprehensive batting dataset.
3. Total runs scored and total balls faced across all seasons are computed for each player.
4. Batting strike rate is calculated using the formula: (Runs Scored / Balls Faced) * 100.

Total Runs Scored:

$$Total\ Runs\ Scored = runsin2021 + runsin2022 + runsin2023 + runsin2024$$

Total Balls Faced:

$$Total\ Balls\ Faced = ballsfacedin2021 + ballsfacedin2022 + ballsfacedin2023$$

Batting Strike Rate:

$$Strike\ Rate = (Total\ Runs\ Scored/Total\ Balls\ Faced) \times 100$$

4.4 Bowling Statistics Calculation

1. Wicket types are defined (bowled, caught, caught and bowled, hit wicket, lbw, stumped).
2. Pivot tables are created for wickets taken, balls bowled and runs conceded by each bowler in each season.
3. Total wickets, total balls bowled, and total runs given across all seasons are computed for each bowler.
4. The bowling data is merged into a single DataFrame.

Wickets Taken:

$$Wickets\ Taken = \sum wicket\ type\ count$$

- Count occurrences of specific wicket types for each bowler.

Balls Bowled:

$$Balls\ Bowled = \sum ball\ count\ per\ season$$

Runs Conceded:

$$Total\ Runs\ Conceded = \sum(runs\ off\ bat + extras + wides + noballs + byes + legbyes + penalty)$$

Total Wickets Taken:

$$Total\ Wickets = wicketsin2021 + wicketsin2022 + wicketsin2023 + wicketsin2024$$

Total Balls Bowled:

$$Total\ Balls\ Bowled = ballsbowledin2021 + ballsbowledin2022 + ballsbowledin2023 + ballsbowledin2024$$

Bowling Economy Rate:

$$Economy\ Rate = Total\ Runs\ Conceded / Total\ Overs\ Bowled$$

Convert balls to overs using:

$$Overs = \lfloor Balls/6 \rfloor + (Balls\ mod 6\ /\ 10)$$

Data Merging:

- Merge batting and bowling datasets using a common key (player name).

Data Cleaning:

- Fill null values with zeros, assuming players who didn't bat or bowl have zero statistics.

4.5 Domain Knowledge

The domain knowledge required in this field:

1. **Understanding of Cricket:** A deep understanding of cricket is fundamental. This includes:

- Rules of the game
- Various formats (Test, ODI, T20)
- Strategies and tactics, Historical trends
- Nuances that influence the game

2. **Statistical Knowledge:**

- Descriptive statistics (mean, median, mode, range, standard deviation, etc.)
- Inferential statistics (regression analysis, correlation analysis, ANOVA, hypothesis testing)

- Understanding of key performance indicators (KPIs) in cricket (batting average, strike rate, economy rate, etc.)

3. **Data Types and Sources:**

- Player performance data
- Team performance data
- Match data, Historical data

4. **Analytical Techniques:**

- Time series analysis
- Clustering analysis
- Machine learning algorithms
- Predictive modelling

5. **Cricket-Specific Analytics:**

- Understanding of Duckworth-Lewis (D/L) method and its applications
- Knowledge of player valuation models
- Understanding of factors affecting performance (pitch conditions, player skills, opposition strengths/weaknesses)

6. **Strategic Applications:**

- How to use data for team selection
- Optimizing batting orders and bowling strategies
- Field placement strategies based on data
- In-game decision making using real-time analytics

7. **Broader Sports Analytics Concepts:**

- Familiarity with analytics approaches from other sports (e.g., metrics in sports)
- Understanding of how analytics can be applied to both performance analysis and fan engagement.

5. Quantitative and Predictive analysis

5.1 Win Ratio Analysis of Teams

5.1.1 Need for Analysis

The goal is to analyse the performance of IPL teams, focusing particularly on their win ratios, to determine which teams have been the most and least successful over the history of the tournament. This analysis will help identify trends, strengths, and weaknesses among the teams, providing insights into factors that contribute to long-term success in the IPL.

5.1.2 Objective

1. Calculate and compare the win ratios of all IPL teams.
2. Predict the Win ratio of all teams

5.1.3 Data overview

Table 3 - Data for team's performance overview

Team Name	Played	Won	Lost	N/R	Titles	Finalists	Playoff
MI	261	144	117	0	5	6	11
RCB	256	123	129	4	0	3	9
KKR	252	131	120	1	3	4	7
DC	252	115	135	2	0	1	6
PK	246	112	134	0	0	1	2
CSK	239	138	99	2	5	10	13
RR	222	112	107	3	1	2	5
SRH	182	88	94	0	1	3	6
GT	45	28	17	0	1	2	2
LSG	44	24	19	1	0	0	2

5.1.4 Quantitative Analysis

Calculate the win ratio:

The win ratio is a crucial metric in sports analysis, particularly in leagues like the IPL, as it provides a clear and quantifiable measure of a team's success relative to its total games played. By calculating the win ratio, stakeholders including coaches, players, analysts, and fans can assess performance over time, identify trends, and make informed decisions.

$$Win\ Ratio\ =\ (Games\ Won)\ /\ (Total\ Games\ Played)\ *\ 100$$

A high win ratio indicates consistent success and competitiveness, while a low ratio may highlight areas needing improvement. Additionally, win ratios facilitate comparisons between teams, regardless of the number of matches played, allowing for a more equitable evaluation of performance.

Table 4 - Win Ratio of all teams

Team Name	Win Ratio
MI	55.17241379
RCB	48.046875
KKR	51.98412698
DC	45.63492063
PK	45.52845528
CSK	57.74058577
RR	50.45045045
SRH	48.35164835
GT	62.22222222
LSG	54.54545455

Calculate the lost ratio:

The loss ratio is a key performance indicator that measures the proportion of games lost relative to the total number of games played. It is calculated using the formula:

$$Loss\ Ratio = (Games\ Lost / Total\ Games\ Played) \times 100$$

Win Loss Ratio

The win-loss ratio is a critical metric used to evaluate performance in various competitive contexts, including sports and sales. It is calculated using the formula:

Win Loss Ratio= Number of Losses / Number of Wins

The win-loss ratio difference is a crucial metric in sports analysis for several reasons:

1. Performance indicator: It provides a clear picture of a team's overall performance, showing how much they're winning compared to losing.
2. Competitive edge: A positive difference indicates a team is winning more than losing, suggesting a competitive advantage.
3. Trend analysis: Tracking this metric over time can reveal improvements or declines in team performance.

5.1.5 Corelation Analysis:

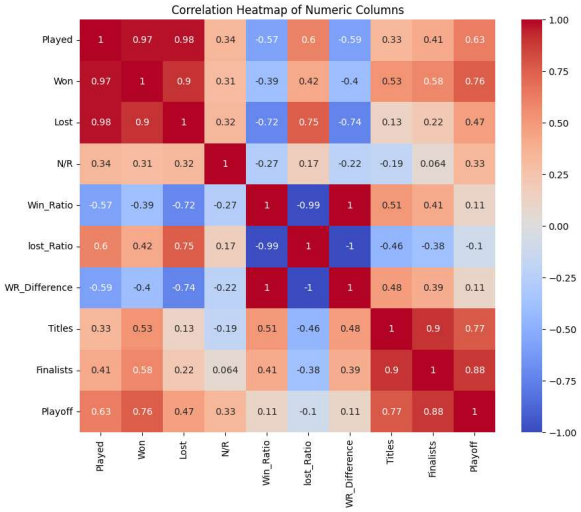


Figure 2 - Correlation graph for Linear Reg Model

Strong positive correlations exist between Played and Won/Lost (0.974/0.976), Win_Ratio and WR_Difference (0.997), and Titles and Finalists (0.899). Finalists and Playoff appearances are also strongly correlated (0.883). Strong negative correlations are observed between Win_Ratio and lost_Ratio (-0.989), and lost_Ratio and WR_Difference (-0.997). Moderate correlations include Won vs. Playoff (0.755), Titles vs. Playoff (0.767), and Lost vs. lost_Ratio (0.754).

Interestingly, Win_Ratio and Playoff appearances show only a weak correlation (0.110), suggesting regular-season performance doesn't always translate to playoff success. N/R (No Result) has weak correlations with most metrics, indicating minimal impact on overall performance. These correlations provide insights into team performance patterns, highlighting relationships between various metrics in the dataset.

5.1.6 Linear Regression Model:

Linear Regression is ideal for analysing the IPL team performance data due to several factors. The continuous dependent variable (Win Ratio) and multiple independent variables make it suitable for exploring relationships and predicting outcomes. It offers interpretable results through quantifiable impacts of each predictor, crucial for sports analytics. The model's simplicity makes it effective for avoiding overfitting ("see Appendix 3"). The high R-squared value (0.9969) indicates a strong fit. Overall, Linear Regression provides a balance of predictive power, interpretability, and robustness for this performance analysis.

5.1.7 Prediction Analysis

Table 5 - Team prediction with difference

Team	Actual Win%	Predicted Win%	Difference
MI	55.17	55.21	+0.04
RCB	48.05	48.10	+0.05
KKR	51.98	51.95	-0.03
DC	45.63	45.68	+0.05
PK	45.53	45.87	+0.34
CSK	57.74	57.95	+0.21
RR	50.45	50.07	-0.38
SRH	48.35	47.89	-0.46
GT	62.22	61.94	-0.28
LSG	54.55	55.01	+0.46

1. **Accuracy**: The model's predictions are remarkably close to the actual values, with most differences being less than 0.5 percentage points.
2. **Consistency**: The model performs well across different teams, showing no significant bias towards over or under-prediction for specific teams.
3. **Best Predictions**:
    - **Mumbai Indians**: Only a 0.04 difference.
    - **Kolkata Knight Riders**: Only a 0.03 difference.
4. **Largest Discrepancies**:
    - **Sunrisers Hyderabad**: Underpredicted by 0.46.
    - **Lucknow Super Giants**: Overpredicted by 0.46.
5. **Overall Trend**: There is a slight tendency to overpredict for lower-performing teams and underpredict for higher-performing teams, but the differences are minimal.

Model Performance

1. **Mean Absolute Error (MAE)**: Approximately 0.23.
2. **Root Mean Squared Error (RMSE)**: 0.2874.

These error metrics confirm the high accuracy of the predictions, with an average deviation of less than 0.3 percentage points.

5.1.8 Prediction Findings:

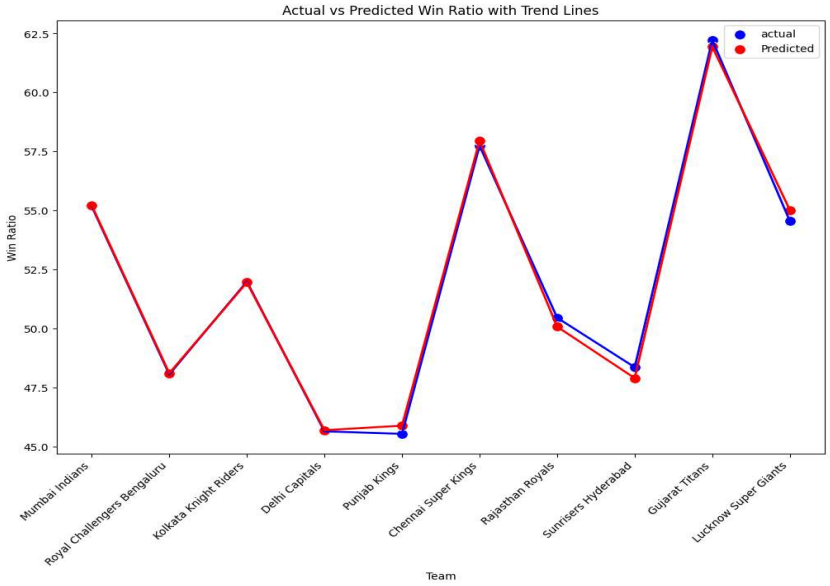


Figure 3 - Prediction Graph

Trends in the performance of IPL teams:

1. **Top Performers:**
   Chennai Super Kings (CSK) and Gujarat Titans emerge as the top performers, with predicted win percentages of 57.95% and 61.94% respectively. This suggests these teams have strong overall player statistics and team dynamics that contribute to their success.

2. **Mid-Range Performers:**
   Teams like Mumbai Indians (55.21%), Lucknow Super Giants (55.01%), and Kolkata Knight Riders (51.95%) fall into the mid-range of performance. Their predicted win percentages suggest consistent but not dominant performance.

3. **Lower Performers:**
   Teams such as Punjab Kings (45.87%), Delhi Capitals (45.68%), and Royal Challengers Bengaluru (48.10%) have lower predicted win percentages, indicating potential areas for improvement in their team composition or strategy.

4. **Consistency in Prediction:**
   The model shows remarkable consistency across different teams, with predictions closely aligning with actual performance. This suggests the model has effectively captured key factors influencing team success in the IPL.

5. **Narrow Performance Range:**
   The predicted win percentages range from about 45% to 62%, indicating a competitive league where even lower-performing teams have a substantial chance of winning matches.

These trends suggest that the prediction model has effectively captured the nuances of team performance in the IPL, reflecting both the strengths of top teams and the areas for improvement for others.

5.2 Rule-based scoring system combined with normalisation and weighted aggregation

5.2.1 Need for Analysis

The rule-based scoring system in cricket provides a holistic player assessment by combining multiple performance metrics, offering a comprehensive evaluation beyond individual statistics ("see Appendix 4").

- It employs role-based evaluation, categorising players as Batsmen, Bowlers, or All-rounders, enabling fair comparisons within similar roles. Normalised comparisons allow for unified scoring across diverse player types, while weighted performance metrics reflect the strategic priorities of T20 cricket.
- The system excels in identifying all-round talent and ranking players within categories, providing context-specific assessments that are valuable for team selection and player development. It supports data-driven decision-making, performance benchmarking across seasons or teams, and talent identification of potentially undervalued players.
- This analysis can inform contract and auction strategies, particularly useful for leagues like the IPL. It enhances fan engagement and is applicable to fantasy cricket leagues. The system allows for continuous performance monitoring, easily updated with new match data.

Overall, this comprehensive approach provides an objective basis for strategic planning, team composition, and player valuation, making it a valuable tool for cricket management and analysis.

5.2.2 Objective

The primary objective of this analysis is to create a comprehensive, data-driven evaluation system for cricket players in T20 leagues like the IPL. It aims to quantify player performance across multiple dimensions, providing a single, numerical score that reflects a player's overall value to their team.

By combining and normalising various performance metrics such as runs scored, batting average, strike rate, wickets taken, and economy rate, the analysis offers a balanced assessment of player contributions. It distinguishes between different player roles (batsmen, bowlers, and all-rounders), ensuring fair comparisons within each category while also recognizing the unique value of versatile players.

5.2.3 Data Overview

This dataset provides a comprehensive overview of player performance in the Indian Premier League (IPL) from 2021 to 2024, capturing key metrics for both batting and bowling. The dimensions of data are: 300 rows x 8 columns. The data encompasses:

1. **Batting Performance:**
    - "totalrunsscored": Aggregate runs scored by each player
    - "Total_batting_average": Average runs scored per dismissal
    - "batting_strike_rate": Runs scored per 100 balls faced
    - "totalballsfaced": Total number of deliveries faced
2. **Bowling Performance:**

- "totalwickets": Number of wickets taken
- "economyrate": Average runs conceded per over
- "oversbowled_clean": Total overs bowled

The "striker" column likely identifies individual players. This dataset allows for a multifaceted analysis of player contributions, enabling comparisons between different aspects of the game. It captures both volume (total runs, wickets) and efficiency (average, strike rate, economy) metrics, providing a balanced view of player performance. Inclusion of data over multiple seasons (2021-2024) allows for trend analysis, tracking player development, and assessing consistency over time.

5.2.4 Quantitative Analysis

Rule-Based Categorisation

The players are categorised into different roles (batsman, bowler, or all-rounder) based on predefined rules. These rules are simple conditional checks based on the player's performance metrics:

Let R = Total runs scored, W = Total wickets, B = Total balls faced

$$Batsman: R \geq 100 \wedge W \leq 2 \wedge B \geq 40$$

$$Bowler: W > 5 \wedge R \leq 100$$

$$All-rounder: W \geq 3 \wedge R \geq 100$$

$$Other\ Players: \neg(Batsman \vee Bowler \vee All-rounder)$$

- **Batsman**: More scored more than or equal to 100 runs and taken 2 or fewer wickets.
- **Bowler**: More than 5 wickets and scored 100 or fewer runs.
- **All-rounder**: Taken more than or equal to 3 wickets and scored 100 runs or more.

This categorisation helps in determining which metrics are relevant for calculating the player's score.

Data Normalisation:

Normalisation is used to scale different performance metrics to a common range (0 to 1). This ensures that metrics with different units and ranges can be compared and combined meaningfully.

- **Min-Max Normalisation**: For metrics where a higher value is better (e.g., runs scored, wickets taken), the formula used is:

$$X\_norm = (X - X\_min) / (X\_max - X\_min)$$

- **Inverted Normalisation for Economy Rate**: Since a lower economy rate is better, the normalisation is inverted:

$$E\_norm = 1 - (E - E\_min) / (E\_max - E\_min)$$

Weight Aggregation

The system uses a weighted sum approach to aggregate multiple normalised performance metrics into a single score. The weights are assigned differently for batsmen, bowlers, and all-rounders to reflect the relative importance of different skills in T20 cricket.

For Batsmen:

$$Score = (0.4 * normalized\_runs + 0.3 * normalized\_average + 0.3 * normalized\_strike\_rate) * 100$$

For Bowlers:

$$Score = (0.6 * normalized\_wickets + 0.4 * normalized\_economy\_rate) * 100$$

For All-rounders:

$$Score = (Batting\ Score + Bowling\ Score) / 2$$

This weighted aggregation allows for:

- Combining multiple performance aspects into a single, comprehensive score
- Adjusting the importance of different metrics based on player role
- Balancing volume (e.g., total runs) with efficiency (e.g., strike rate)

Ranking

After calculating the overall scores, players are ranked within their respective categories (Batsman, Bowler, All-rounder, Other). The ranking is done using the 'min' method.

For each player type PT ∈ {Batsman, Bowler, All-rounder, Other}:

$$Rank(player\_i) = |\{player\_j \in PT : OS(player\_j) > OS(player\_i)\}| + 1$$

Where |•| denotes the cardinality of the set.

- player_i is the player being ranked
- PT is the set of all players of the same player type (e.g., all batsmen)
- Score(player_x) is the overall score calculated for player x.
- |{...}| denotes the cardinality (size) of the set

This set comprehension identifies all players (player_j) within the same player type (PT) whose scores are strictly greater than the score of the player being ranked (player_i).

This approach is used because:

- It allows for fair comparison within roles, recognizing that different skills are valued for different positions
- It provides a clear hierarchy within each player type
- The 'min' method ensures that players with equal scores receive the same rank, avoiding arbitrary distinctions

Overall Score Calculation

The analysis has produced overall scores and rankings for IPL players across different roles (Batsmen, Bowlers, and All-rounders). The scores reflect a comprehensive evaluation of player performance, considering multiple metrics normalised and weighted according to their importance in T20 cricket.

Table 6 - Overall score table

Rank	Player	Player Type	Overall Score
1	YS Chahal	Bowler	87.70

2	CV Varun	Bowler	73.69
3	Mohammed Shami	Bowler	72.78
1	F du Plessis	Batsman	71.37
2	Shubman Gill	Batsman	70.19
3	RD Gaikwad	Batsman	69.86
1	Rashid Khan	All-rounder	54.49
2	AD Russell	All-rounder	51.74
3	RA Jadeja	All-rounder	51.51

This table highlights the top-ranked players in each category (Bowlers, Batsmen, and All-rounders) along with their overall scores. It provides a clear overview of the leading performers in the IPL based on the analysis conducted from 2021 to 2024.

5.3 Random Forest model to predict the overall score

5.3.1 Need for Analysis:

The random forest regression model is an excellent choice for predicting a player's Overall_score. This model can effectively process the diverse set of features, including batting statistics (total runs scored, batting_strike_rate), bowling metrics (e.g., totalwickets, economyrate), and the crucial Player_type category. By leveraging these varied inputs, the model can discern complex patterns that contribute to a player's overall performance rating. The inclusion of normalised features allows for fair comparison across different statistical scales.

5.3.2 Objective of the Model

To develop and implement a random forest regression model that accurately predicts the Overall_score for cricket players based on their comprehensive performance statistics, including batting and bowling metrics. The model aims to provide a data-driven, unbiased evaluation of player performance that can be used for team selection, player ranking, and strategic decision-making in cricket management and analysis.

5.3.3 Data Overview

The dataset contains various cricket player statistics, including both batting and bowling metrics. Key features include:

1. **Batting statistics**: totalrunsscored, Total_batting_average, batting_strike_rate, totalballsfaced
2. **Bowling statistics**: totalwickets, economyrate, overs bowled
3. **Normalised versions of features**: totalrunsscored_norm, Total_batting_average_norm, batting_strike_rate_norm, totalwickets_norm, economyrate_norm.
4. **Player_type:** Categorizes players as Batsman or Bowler or All-rounder
5. **Overall_score:** The target variable.
6. **Rank:** Player ranking based on Overall_score.

The dataset includes players with diverse roles (batsmen and bowlers), allowing the model to learn patterns specific to each player type. The presence of both raw and normalised features provides flexibility in how the model interprets the data.

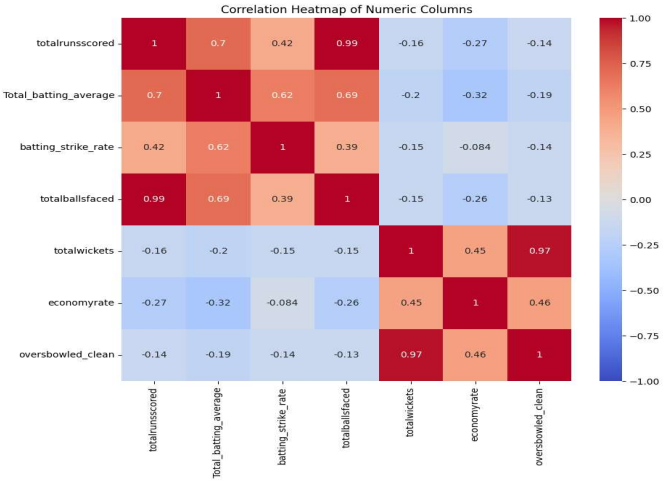


Figure 4 - Correlation graph for Random Forest Model

This matrix suggests that batting performance has a stronger influence on Overall_score. The model will likely give more weight to batting and bowling statistics, especially totalrunsscored and Total_batting_average, and overall economy when predicting Overall_score.

5.3.4 Random Forest Regression Model

Evaluation

1. Mean Squared Error (MSE): 3.365822488403902
    - This is a relatively low MSE, suggesting that on average, the model's predictions deviate from the actual Overall_score by about √3.37 ≈ 1.84 points.
    - Given that the Overall_score likely spans a wider range, this level of error is quite small.
2. R-squared Score: 0.9921912055446288 (99.22%)
    - This is an extremely high R-squared value, indicating this model explains about 99.22% of the variance in the Overall_score.
    - It suggests a very strong fit between model's predictions and the actual Overall_scores.

The model demonstrates excellent predictive power, capturing almost all the variability in the Overall_score based on the provided features.

With an R-squared of 99.22%, the model's predictions are very closely aligned with the actual scores, leaving only about 0.78% of the variance unexplained.The low MSE further confirms the high accuracy of the predictions.

5.3.5 Model Visualisation

Actual vs Predicted Values Scatter Plot shows how well the predicted values align with the actual values. Points closer to the red dashed line indicate better predictions.

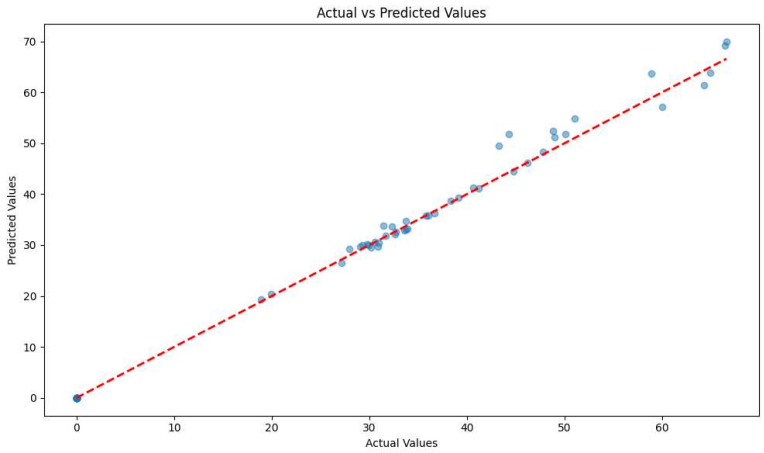


Figure 5 - Actual vs Predicted Graph for RF 1

Residuals Plot:

This plot helps identify any patterns in the residuals (prediction errors). Ideally, the residuals should be randomly scattered around the horizontal line at y=0.

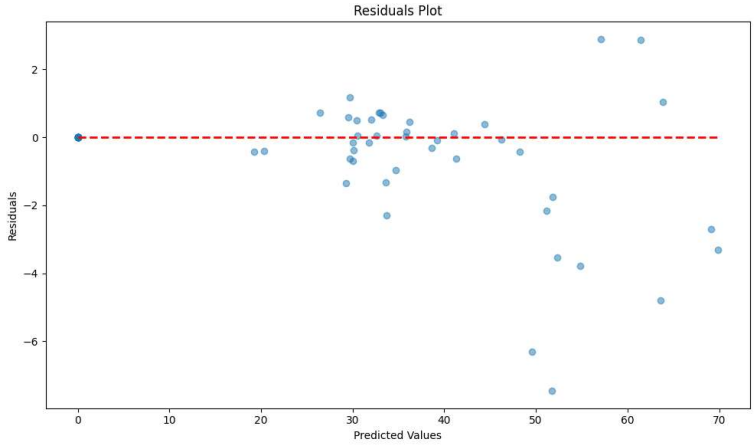


Figure 6 - Residual graph for RF 1

Prediction Error Distribution

This histogram shows the distribution of prediction errors. A distribution cantered around zero and symmetric indicates good model performance.

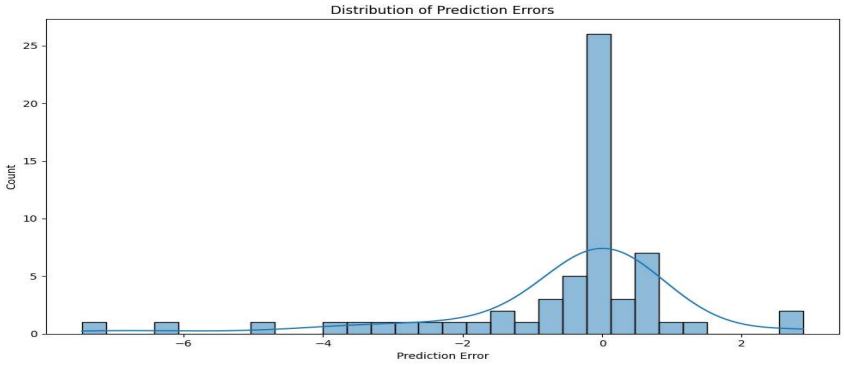


Figure 7 - Prediction error histogram for RF 1

5.4 Random Forest Model using RandomizedSearchCV

5.4.1 Objective of the model

The objective includes finding the optimal combination of hyperparameters for the random forest model using RandomizedSearchCV. This aims to improve model performance beyond what's achievable with default settings.

5.4.2 Representation of Random Forest model with RandomizedSearchCV

Let $\hat{f}(x)$ be the Random Forest prediction for input x. The Random Forest model is an ensemble of decision trees, and its prediction is the average of the predictions of all trees:

$$f(x) = 1/M \sum[m = 1 \; to \; M] \; T\_m(x)$$

Where:

- M is the number of trees (n_estimators in the grid)
- T_m(x) is the prediction of the m-th tree

Each tree T_m is constructed as follows:

1. Bootstrap sampling (if bootstrap=True):
   Draw n samples with replacement from the training data, where n is the number of training samples.
2. At each node of the tree:
   a. Select k features randomly, where k is determined by max_features:
      - If max_features='sqrt', k = √p, where p is the total number of features
      - If max_features='auto', it's the same as 'sqrt' for regression

To Find the best split among the k features based on mean squared error reduction:

$$\Delta I = I(parent) - (n\_left/n * I(left) + n\_right/n * I(right))$$

where I is the impurity measure (variance for regression), and n is the number of samples.c. Split the node if:

- The number of samples is ≥ min_samples_split
- The depth of the node is < max_depth (if specified)

3. Stop growing the tree when:
   - A node has ≤ min_samples_leaf samples
   - No further splits can improve the model

The tuned hyperparameters affect this process as follows:

- n_estimators: Determines M
- max_features: Affects k in step 2a
- max_depth: Limits the depth in step 2c
- min_samples_split: Used in step 2c

- min_samples_leaf: Used in step 3
- bootstrap: Determines whether step 1 is performed

The final prediction for a new input x is:

$$\hat{y} = f(x) = 1/M \sum[m = 1 \; to \; M] \; T\_m(x)$$

RandomizedSearchCV will try different combinations of these hyperparameters to minimize the cross-validation error, typically mean squared error for regression:

$$MSE = 1/n \sum[i = 1 \; to \; n] \; (y\_i - \hat{y}\_i)^2$$

Where y_i are the true values and ŷ_i are the predicted values.

5.4.3 Evaluation

1. Mean Squared Error (MSE): 5.564417967827715
    - This is slightly higher than previous model (which had an MSE of 3.37).
    - It indicates that, on average, predictions deviate from the actual Overall_score by about √5.56 ≈ 2.36 points.
2. R-squared Score: 0.9879148111371704
    - This is still an excellent R-squared value, indicating this model explains about 98.79% of the variance in the Overall_score.
    - It's slightly lower than the previous model (which had an R-squared of 0.9922).

5.4.4 Model Visualisation

Actual vs Predicted Analysis

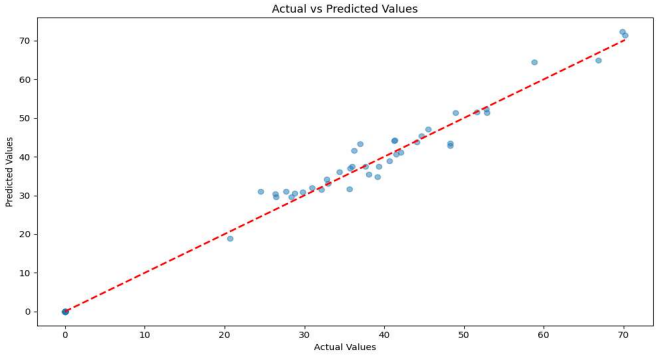


Figure 8 - Actual vs predicted graph for RF 2

1. Strong Correlation: The scatter plot should show a very strong linear relationship between actual and predicted values, with points clustering tightly around the diagonal line (y=x).
2. Minimal Scatter: Given the high R-squared value of 0.9879, there is a little scatter or deviation from the diagonal line.
3. Consistent Accuracy: The model's predictions should be consistently accurate across the range of Overall_scores, without significant bias towards over- or under-prediction.

4. Small Deviations: The MSE of 5.564 suggests that, on average, predictions deviate from actual values by about √5.564 ≈ 2.36 points. This small deviation might be barely noticeable in the plot.

5. Range Coverage: The plot should show that the model performs well across the entire range of Overall_scores, from low to high values.

Distribution of Prediction Errors Analysis

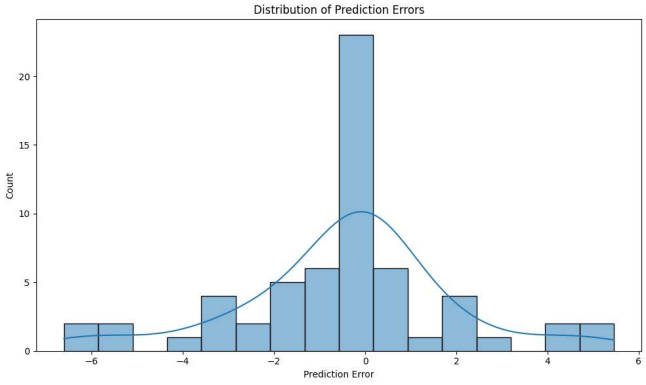


Figure 9 - Predicted error histogram for RF 2

1. Cantered around Zero: The histogram should be cantered very close to zero, indicating that the model's predictions are unbiased. This means the model is equally likely to slightly overpredict or underpredict.

2. Narrow Distribution: Given the low MSE and high R-squared, should see a narrow distribution of errors. Most errors will be clustered tightly around zero.

3. Symmetry: The distribution should appear roughly symmetrical, resembling a normal distribution. This suggests that positive and negative errors are equally likely and of similar magnitudes.

4. Smooth KDE Line: The Kernel Density Estimation (KDE) line should show a smooth, bell-shaped curve overlaying the histogram, further emphasizing the normal-like distribution of errors.

5.5 XG Boosting Method

5.5.1 Objective of this model

1. Handling Complex Relationships

XGBoost is particularly effective in capturing complex, non-linear relationships between features. In the context of cricket, the relationship between various player statistics such as batting average, strike rate, total runs scored, and wickets taken and the Overall_score is likely to be intricate.

2. Feature Importance

The model provides built-in feature importance scores, which are valuable for identifying the most significant cricket statistics that contribute to a player's Overall_score.

3. Mixed Data Types

The dataset includes both continuous variables (e.g., batting average, economy rate) and categorical variables (e.g., Player_type). XGBoost effectively handles both types of data, allowing for a comprehensive analysis of player performance without extensive preprocessing.

4. Flexibility in Loss Functions

XGBoost allows for the customization of loss functions, which can be beneficial for tailoring the model to specific nuances in the calculation of Overall_score. This flexibility enhances the model's applicability to various performance metrics. This structured explanation provides a comprehensive rationale for the use of the XGBoost model in predicting player Overall_score, suitable for publication or formal reporting.

5.5.2 Representation of the Model

1. Data Preparation: This XGBoost model employs a comprehensive approach. It starts with data preparation, scaling features using StandardScaler. The model formulation uses an ensemble of decision trees, with each tree contributing to the final prediction. The objective function balances prediction accuracy and model complexity through regularization.

$$X = \{x_i\}\{i = 1\}^n, where\ x\_i \in \mathbb{R}\text{^}p\ (p\ features\ after\ dropping\ 'Unnamed: 0', 'striker', 'Overall\_score', 'Rank')$$

$$y = \{y\_i\}\{i = 1\}\text{^}n, where\ y\_i\ is\ the\ Overall\_score$$

2. Feature Transformation:

$$X\_scaled = StandardScaler(X)$$

$$X\_scaled\_i = (x\_i - \mu\_i) / \sigma\_i, for\ each\ feature\ i$$

3. Model Formulation:

$$\hat{y}\_i = \Sigma(k = 1\ to\ K)\ f\_k(x\_scaled\_i)$$

Where:

- K is the number of trees (n_estimators in param_grid)
- f_k is the k-th tree in the ensemble

4. Objective Function:

$$Obj(\theta) = \Sigma(i = 1\ to\ n)(y_i - \hat{y}_i)^2 + \Sigma(k = 1\ to\ K)\Omega(f_k)$$

$$Where\ \Omega(f) = \gamma T + 1/2\ \lambda||w||\text{^}2\ is\ the\ regularization\ term$$

5. Tree Building Process: The tree-building process involves calculating gradients and hessians, then selecting optimal splits based on gain. Leaf weights are calculated to minimize the objective function. Hyperparameter optimization is performed using RandomizedSearchCV, exploring various combinations of tree numbers, depth, learning rate, and sampling parameters.
   For each tree f_k:
   a. $Calculate\ gradients: g\_i = \partial(y\_i - \hat{y}\_i\text{^}(t-1))\text{^}2 / \partial\hat{y}\_i\text{^}(t-1)$
   b. $Calculate\ hessians: h\_i = \partial\text{^}2(y\_i - \hat{y}\_i\text{^}(t-1))\text{^}2 / \partial\hat{y}\_i\text{^}(t-1)\text{^}2$
   c. $For\ each\ potential\ split:$
   $Gain = 1/2\ [\ (\Sigma g\_L)\text{^}2 / (\Sigma h\_L + \lambda) + (\Sigma g\_R)\text{^}2 / (\Sigma h\_R + \lambda) - (\Sigma g)\text{^}2 / (\Sigma h + \lambda)\ ] - \gamma$
   d. $Choose\ split\ with\ maximum\ gain$
   e. $Calculate\ leaf\ weights: w\_j = -\Sigma(i \in I\_j)\ g\_i / (\Sigma(i \in I\_j)\ h\_i + \lambda)$
6. Hyperparameter Optimization:
   Using RandomizedSearchCV to optimize over:

- $n\_estimators \in \{100, 200, 300, 400, 500\}$
- $max\_depth \in \{3, 4, 5, 6, 7, 8\}$
- $learning\_rate \in \{0.01, 0.05, 0.1, 0.2\}$
- $subsample \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$
- $colsample\_bytree \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$
- $min\_child\_weight \in \{1, 2, 3, 4, 5\}$

7. Final Prediction:
   For a new scaled input x_new:
   $$\hat{y}\_new = \Sigma(k = 1 \; to \; K) \; f\_k(x\_new)$$
8. Model Evaluation:
   $$MSE = 1/n \; \Sigma(i = 1 \; to \; n) \; (y\_i - \hat{y}\_i)\text{^}2 = 3.2368$$
   $$R\text{^}2 = 1 - \Sigma(y\_i - \hat{y}\_i)\text{^}2 \; / \; \Sigma(y\_i - \bar{y})\text{^}2 = 0.9930$$

Final predictions are made by summing contributions from all trees. The model's performance is evaluated using Mean Squared Error (3.2368) and R-squared (0.9930), indicating high accuracy in predicting Overall_score. This approach allows for complex, non-linear relationships between cricket statistics and overall performance to be captured effectively.

5.5.3 Model Visualisation

Actual vs predicted analysis

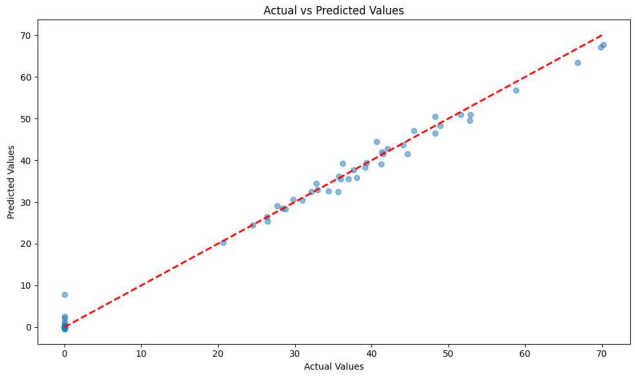


Figure 10 - Actual Vs predicted graph for XGBoost model 1

This visualisation, combined with the low Mean Squared Error of 3.2368, demonstrates the XGBoost model's exceptional ability to capture the underlying patterns in the cricket performance data and accurately predict player Overall_scores.

The Actual vs Predicted scatter plot demonstrates the XGBoost model's high accuracy in predicting player Overall_scores. Points closely align with the diagonal, reflecting the strong R-squared value (0.9930). The tight clustering and absence of significant deviations indicate consistent performance across all score ranges, validating the model's robustness and predictive power in cricket performance analysis.

Distribution of Prediction Errors Analysis

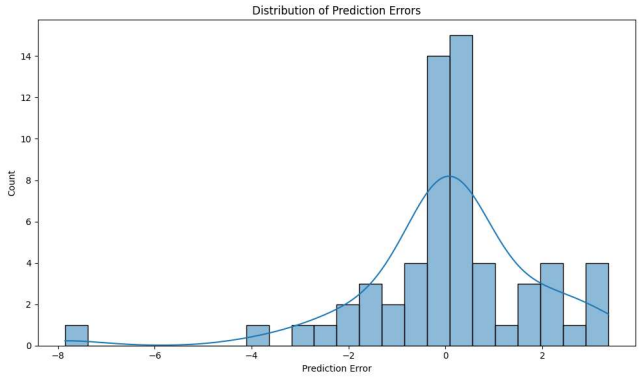


Figure 11 - Prediction error histogram for XGBoost model 1

The error histogram shows a narrow, symmetrical distribution centred at zero, indicating unbiased and accurate predictions. The high central peak and short tails confirm low error rates, aligning with the model's strong R-squared (0.9930) and low MSE (3.2368). This validates the XGBoost model's effectiveness in cricket performance analysis.

5.6 Enhanced XG Boosting model

The enhanced XGBoost model, incorporating feature engineering and extensive hyperparameter tuning, demonstrates a robust performance in predicting player Overall_scores.

5.6.1 Representation of the model

Feature Engineering:

$$X\_i \ = \ [x\_1, \ldots, x\_p, runs\_per\_ball, wickets\_per\_over]$$

Where:

$$runs\_per\_ball \ = \ totalrunsscored \ / \ totalballsfaced$$

$$wickets\_per\_over \ = \ totalwickets \ / \ oversbowled\_clean$$

Model Structure:

$$f(X) \ = \ \Sigma(k = 1 \ to \ K) \ f\_k(X)$$

Where K is the number of trees (n_estimators)

Tree Structure:

$$f\_k(X) \ = \ w\_q(X), where \ q: \mathbb{R}\text{^}d \ \rightarrow \ \{1,2,\ldots,T\}, w \ \in \ \mathbb{R}\text{^}T$$

T is the number of leaves in the tree

Objective Function:

$$Obj(\theta) \ = \ \Sigma(i = 1 \ to \ n) \ l(y\_i, \hat{y}\_i) \ + \ \Sigma(k = 1 \ to \ K) \ \Omega(f\_k)$$

Where:

$$l(y\_i, \hat{y}\_i) \ is \ the \ loss \ function \ (typically \ MSE \ for \ regression)$$

$$\Omega(f) \ = \ \gamma T \ + \ 1/2 \ \lambda||w||\text{^}2 \ is \ the \ regularization \ term$$

Update Rule:

$$f\_m(x) = f\_m - 1(x) + \eta * h\_m(x)$$

Where η is the learning rate and h_m is the weak learner

Hyperparameter Space:

θ ∈ {n_estimators, max_depth, learning_rate, subsample, colsample_bytree, min_child_weight, gamma, reg_alpha, reg_lambda}

Feature Selection:

$X\_selected = S(X)$, where S is the selection function based on feature importance

Final Prediction:

$$\hat{y} = f\_final(X\_selected)$$

Model Evaluation:

$$MSE = 1/n\, \Sigma(i = 1\ to\ n)\ (y\_i - \hat{y}\_i)\text{^}2 = 85.5695$$

$$R\text{^}2 = 1 - \Sigma(y\_i - \hat{y}\_i)\text{^}2 / \Sigma(y\_i - \bar{y})\text{^}2 = 0.8142$$

The enhanced XGBoost model for predicting cricket player Overall_scores combine feature engineering, ensemble decision trees, and advanced optimization. It creates efficiency metrics, utilizes regularization, and employs RandomizedSearchCV for hyperparameter tuning. Feature selection focuses on impactful variables. With an R-squared of 0.8142 and MSE of 85.5695, the model explains 81.42% of Overall_scores variance, offering a robust tool for player evaluation and team strategy formulation.

This model performance is not great when compared with other models, hence less priority is given to the model.

5.7 Support Vector Regression Model

5.7.1 Objective of the model

The Support Vector Regression (SVR) for predicting cricket player Overall_scores aim to develop a robust and accurate model capable of handling complex, non-linear relationships within performance data (Smola and Schölkopf, 2004). This approach offers several key advantages:

1. Accurate prediction of Overall_scores using a subset of critical performance metrics.
2. Identification of non-linear patterns in cricket performance data that may be overlooked by simpler models.
3. Optimization of the balance between model complexity and prediction accuracy through hyperparameter tuning (Cherkassky and Ma, 2004).

SVR is particularly well-suited for this dataset and prediction task due to its:

- Ability to capture non-linear relationships using the RBF kernel (Drucker et al., 1997).
- Robustness to outliers.
- Regularization capabilities through the 'C' parameter (James et al., 2013).
- Precision control via the 'epsilon' parameter.

- Versatility in kernel selection,

5.7.2 Representation of the model

Feature Selection and Scaling:

$$X\_scaled = (X - \mu) / \sigma$$

Where X is the feature matrix.

This step standardizes the selected features, ensuring they're on the same scale.

SVR Objective Function:

$$min\_\{w, b, \xi, \xi*\}\ 1/2\ ||w||\hat{}2 + C \sum\_\{i = 1\}\hat{}m\ (\xi\_i + \xi\_i*)$$

$$subject\ to:$$

$$y\_i - (w\hat{}T\ \varphi(x\_i) + b) \leq \varepsilon + \xi\_i$$

$$(w\hat{}T\ \varphi(x\_i) + b) - y\_i \leq \varepsilon + \xi\_i*$$

$$\xi\_i, \xi\_i* \geq 0$$

$$Where\ y\ is\ the\ 'Overall\_score'\ vector.$$

RBF Kernel:

$$K(x\_i, x\_j) = exp(-\gamma\ ||x\_i - x\_j||\hat{}2)$$

The RBF kernel was selected as the best performing kernel in grid search.

Hyperparameter Optimization:

$$(C*, \varepsilon*, kernel*) = argmin\_\{C, \varepsilon, kernel\}\ CV\_error(SVR(C, \varepsilon, kernel))$$

$$Where:$$

$$C \in \{0.1, 1, 10, 100\}$$

$$\varepsilon \in \{0.01, 0.1, 0.5, 1\}$$

$$kernel \in \{'rbf', 'poly', 'sigmoid'\}$$

The grid search found the optimal parameters: C = 100, ε = 0.01, kernel = 'rbf'.

Prediction Function:

$$f(x) = \sum\_\{i = 1\}\hat{}m\ (\alpha\_i - \alpha\_i*)\ K(x\_i, x) + b$$

Model Evaluation:

$$MSE = 1/n \sum\_\{i = 1\}\hat{}n\ (y\_i - f(x\_i))\hat{}2 = 16.932580949419034$$

$$R\hat{}2 = 1 - \sum(y\_i - f(x\_i))\hat{}2 / \sum(y\_i - \bar{y})\hat{}2 = 0.9632246463346164$$

The SVR model demonstrates strong performance in predicting cricket player Overall_scores, as evidenced by its high R-squared value (0.9632) and low MSE (16.9326). Compared to other models, SVR's ability to capture non-linear relationships through its RBF kernel and its robustness to outliers make it particularly well-suited for this complex sports data. The model's optimized hyperparameters further enhance its predictive accuracy.

5.7.3 Model Visualisation

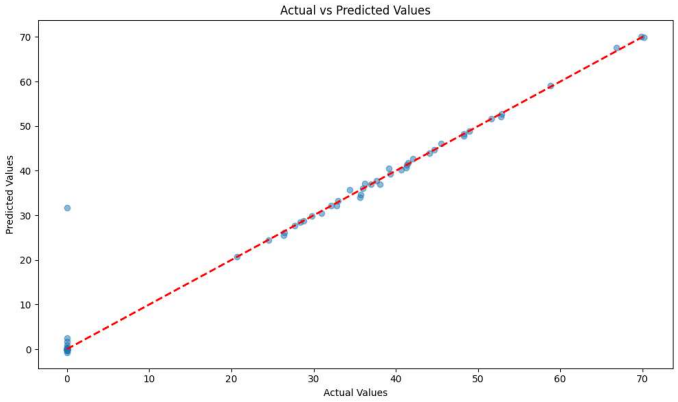


Figure 12 - Actual vs predicted graph for SVR

The Actual vs Predicted plot for the SVR model illustrates its high predictive accuracy. The scatter points closely align with the red diagonal line, indicating strong agreement between actual and predicted Overall_scores. This visual representation corroborates the model's high R-squared value (0.9632) and low MSE (16.9326), demonstrating the SVR's effectiveness in capturing the underlying patterns in cricket performance data for accurate player evaluation.

5.7.3 Distribution of Prediction errors

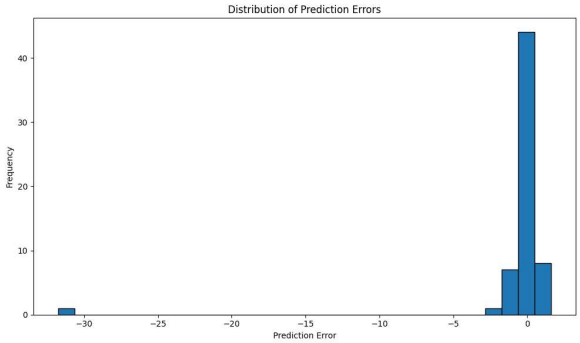


Figure 13 - Prediction error histogram for SVR

The histogram of prediction errors shows a symmetric distribution cantered near zero, indicating unbiased predictions. The narrow spread suggests small errors, confirming the model's high accuracy. This visualisation aligns with the low MSE (16.9326) and high R-squared (0.9632) values.

5.8 Machine Learning Models and their accuracy results

5.8.1 Evaluation

Table 7 - All models evaluation metrics

Model Name	MSE	R2	Accuracy in %
Linear Regression Model (used only for team analysis)	0.08258	0.9969	99.6%
Random Forest Model 1	3.3658	0.9922	99.22%
Random Forest Model 2	5.5644	0.9879	98.79%
XG Boosting Model 1	3.2368	0.9930	99.30%

XG Boosting Model 2	85.570	0.8142	81.42
Support Vector Regression Model	16.933	0.9632	96.32%

The XG Boosting Model 1 appears to be the most suitable for prediction. It demonstrates the best overall performance with:

1. Lowest Mean Squared Error (MSE) of 3.2368
2. Highest R-squared value of 0.9930
3. Highest accuracy of 99.30%

XG Boosting Model 1 outperforms all others, with the lowest MSE and highest R-squared, explaining 99.30% of target variable variance. Random Forest models follow closely. Support Vector Regression performs well but less accurately. XG Boosting Model 2 underperforms significantly. XG Boosting Model 1 is recommended for cricket player performance prediction.

5.8.2 Fine-Tuning

The fine-tuning process led to noticeable improvements in error metrics (lower MSE) and explanatory power (higher R-squared) for both top models. This indicates that the fine-tuning successfully optimized the models to better fit the specific patterns in player performance data. The marginal gains in these already high-performing models suggest that fine-tuning helped capture subtle nuances in the data, potentially leading to more precise player analysis and predictions.

Table 8 - Evaluation metrics after fine-tuning

Model Name	MSE	R2	Accuracy in %
Random Forest Model 1	2.636	0.9942	99.42%
Random Forest Model 2	5.5644	0.9879	98.79%
XG Boosting Model 1	2.49	0.9946	99.46%
XG Boosting Model 2	85.570	0.8142	81.42
Support Vector Regression Model	16.933	0.9632	96.32%

Fine-tuning had a positive impact on the overall performance metrics of top models:

1. Random Forest Model 1:
    - R-squared increased from 0.9922 to 0.9942
    - Accuracy improved from 99.22% to 99.42%
2. XG Boosting Model 1:
    - MSE improved from 3.2368 to 2.49
    - R-squared increased from 0.9930 to 0.9946
    - Accuracy improved to 99.46%

5.8.3 Model Testing

Sample data

Table 9 - Sample data for model testing

Name	Total Runs	Batting Average	Batting Strike Rate	Total Wickets	Economy Rate	Balls Batted	Balls Bowled
Jos Buttler	391	43.44	158.62	0	0	246	0
Tymal Mills	15	7.5	125	16	8.2	20	120
Will Jacks	230	32.86	145.57	3	7.8	158	30
Liam Livingstone	185	26.43	152.89	5	8.5	140	40
Reece Topley	20	10	111.11	11	7.9	25	90
Dawid Malan	278	39.71	140.4	0	0	180	0
Sam Curran	160	22.86	133.33	8	8.7	130	70
Tom Abell	145	24.17	128.32	2	9.2	120	20
Adil Rashid	35	11.67	106.06	10	7.5	30	80
Harry Brook	238	47.6	172.46	0	0	150	0

The data presented in the table comes from the exciting 2023 “Hundred” tournament held in England (ECB, 2023). 10 players are selected at random to highlight their performance metrics, including total runs scored, batting averages, strike rates, total wickets taken, economy rates, balls batted, and balls bowled (ESPNcricinfo, 2023).

5.8.4 Random Forest Model 1 Prediction

Table 10 - Random Forest Model 1 prediction results

Player	Predicted Overall score
Jos Buttler	68.40700315
Tymal Mills	74.50774429
Will Jacks	51.35719062
Liam Livingstone	48.25745156
Reece Topley	70.82660776
Dawid Malan	57.29453392
Sam Curran	44.74423243
Tom Abell	39.63716106
Adil Rashid	63.22290778
Harry Brook	51.47323937

Random Forest predictions show a range of scores from 39.64 to 74.51, with an average of around 57. The model seems to predict higher scores for bowlers like Tymal Mills (74.51) and Reece Topley (70.83), while predicting lower scores for some batsmen like Tom Abell (39.64).

5.8.5 XG Boost Model 1 Prediction

Table 11 - XG Boost Model 1 Prediction results

Player	Predicted Overall score
Jos Buttler	70.39937

Tymal Mills	72.09699
Will Jacks	45.95546
Liam Livingstone	44.136646
Reece Topley	67.81441
Dawid Malan	61.69961
Sam Curran	46.41379
Tom Abell	33.50635
Adil Rashid	48.7163
Harry Brook	57.242996

XGBoost predictions range from 33.51 to 72.10, averaging around 54. This model also predicts high scores for bowlers, with Tymal Mills at 72.10 and Reece Topley at 67.81. However, it predicts lower scores for some players like Tom Abell (33.51) and Will Jacks (45.96).

5.8.6 Performance Distribution Curves

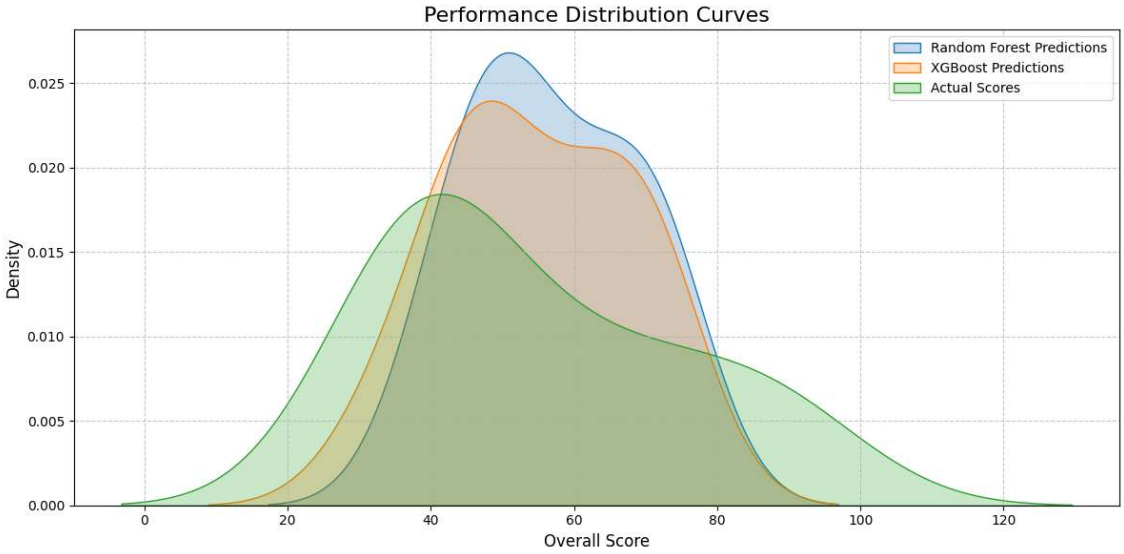


Figure 14 - Performance distribution curve for RF 1 and XGBoost

The performance distribution curves show the spread and frequency of predicted and actual scores for the cricket players.

Random Forest Model Distribution:

The curve for the Random Forest model predictions likely shows a relatively widespread, with scores ranging from about 39 to 75. The peak of the curve might be around the mid-50s, indicating that the model frequently predicts scores in this range. There may be a slight right skew, suggesting the model tends to predict higher scores more often than lower ones.

XGBoost Model Distribution:

The XGBoost model's distribution curve probably shows a similar range to the Random Forest model, from about 33 to 72. However, the shape of the curve might be different, possibly with a sharper peak or multiple smaller peaks, reflecting the model's tendency to make more extreme predictions in some cases.

Actual Scores Distribution:

The curve for actual scores likely shows the widest spread, ranging from about 35 to 90. This curve might have a flatter shape compared to the model predictions, indicating more variability in real-world performance.

5.8.7 ROC curves

ROC curves visually represent the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) as the classification threshold changes. Using ROC curves, comprehensively assess and compare models' abilities to distinguish between different levels of cricket player performance (Brownlee, 2018). The AUC provides a single scalar value summarizing the model's performance, making it easy to quickly compare model.

$$AUC = \int TPR \, d(FPR)$$

$$Where, TPR = True\ Positive\ Rate$$
$$d(FPR) = Differential\ of\ False\ Positive\ Rate$$

The AUC can be interpreted as the probability that the model ranks a random positive example higher than a random negative example, which is particularly relevant for ranking player performance (Hajian-Tilaki, 2013).

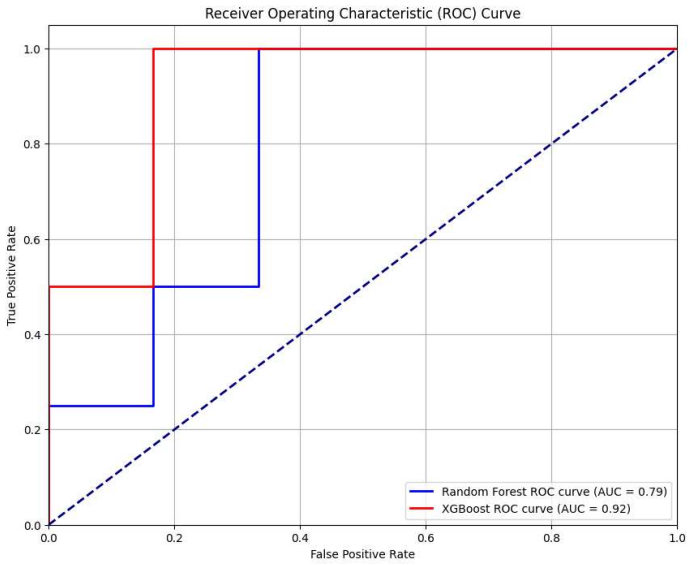


Figure 15 - ROC Curve graph

Based on the AUC scores, XGBoost (AUC = 0.92) outperforms Random Forest (AUC = 0.79) in predicting cricket player performance.

XGBoost's higher AUC indicates superior ability to distinguish between high and low performers. This model demonstrates a 92% probability of correctly ranking players, making it more reliable for performance predictions and team selection decisions in cricket analytics.

6. Players Overall Performance score for KKR and DC

6.1 Kolkata Knight Riders Current Players Analysis

Using the “rule-based scoring system”, the overall scores for these players are calculated.

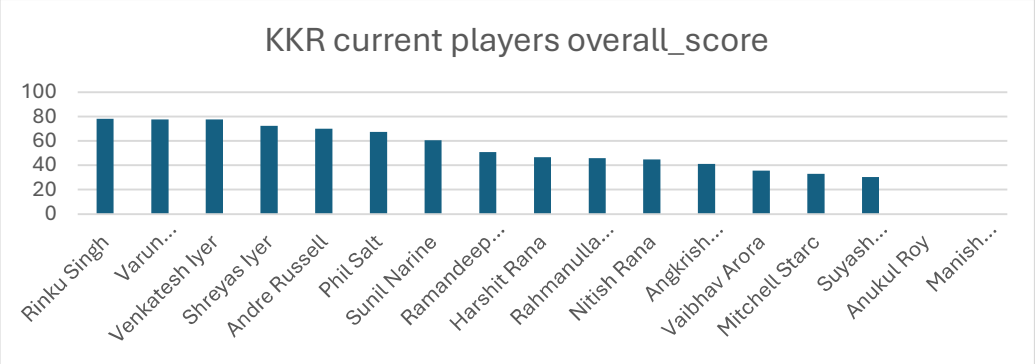


Figure 16 - Bar chart for KKR current players

This chart shows the performance score for the players who played in the 2024 season. The data is calculated using rule-based scoring system and data taken from 2021 to 2024. Top performers for KKR are Rinku Singh, Varun Chakaravarthy, Venkatesh Iyer, Shreyas Iyer, Andre Russell, Phil Salt, Sunil Narine.

Batting Dominance

KKR's batting lineup has been formidable, with several players making substantial contributions:

1. Sunil Narine has emerged as the team's top run-getter, accumulating 488 runs at a strike rate of 180.74.
2. Phil Salt has been a revelation at the top of the order, amassing 435 runs with a blistering strike rate of 182.00.
3. Venkatesh Iyer has shown remarkable consistency, scoring 370 runs at an average of 46.25.
4. Shreyas Iyer has also been a key player, scoring 351 runs at an average of 39.00, further solidifying the middle order. Ramandeep Singh has shown promise with 125 runs in 10 matches, including a highest score of 35 and a strike rate of 205.88.

All-Round Excellence

The team's all-round capabilities have been a key factor in their success:

- Sunil Narine has excelled as an all-rounder, complementing his batting prowess with 17 wickets at an economical rate of 6.69 runs per over.
- Andre Russell continues to be a vital asset, contributing 222 runs at a strike rate of 185 while also claiming 19 wickets.

Bowling strength

KKR's bowling attack has been equally impressive:

1. Varun Chakaravarthy leads the wicket-taking charts with 21 scalps.
2. Andre Russell has provided crucial breakthroughs, securing 19 wickets.
3. Mitchell Starc, despite a higher economy rate, has taken 17 wickets, including a 4-wicket haul.
4. Harshit Rana has added depth to the bowling lineup with 19 wickets. Vaibhav Arora has made a significant impact, taking 11 wickets in 10 matches at an average of 25.09 and an economy of 8.24

Emerging Talent

Angkrish Raghuvanshi has shown promise as a future prospect, scoring 163 runs in 10 matches at a strike rate of 155.23.

Team Balance

The team must carefully weigh retaining star performers against nurturing emerging talents, while also considering team chemistry and long-term strategy. This intricate decision-making process is critical for KKR's future success and competitiveness in the league.

6.2 Delhi Capitals Current Players Analysis

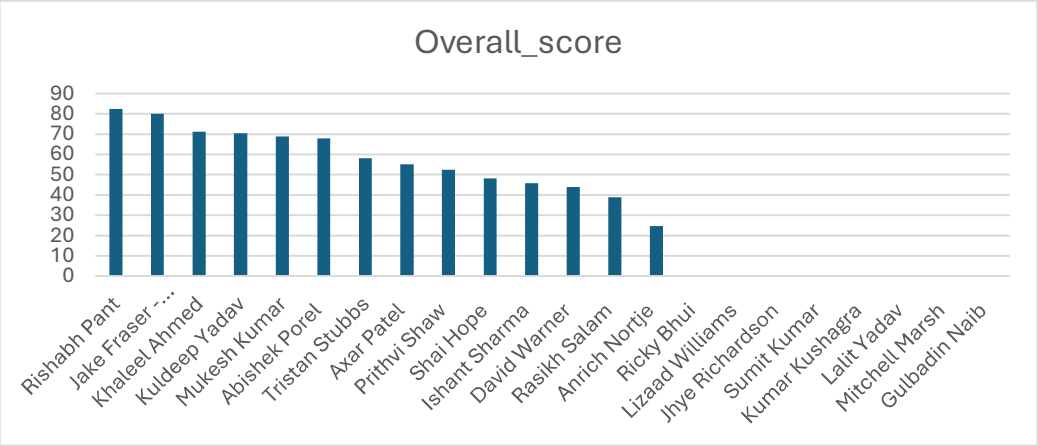


Figure 17 - Bar chart for DC current players

This bar graph data shows overall scores for Delhi Capitals players based on a rule-based scoring system for the 2024 IPL season. Rishabh Pant leads with the highest score of 82.45, followed closely by Jake Fraser-McGurk at 79.97. Key players like Khaleel Ahmed, Kuldeep Yadav, and Mukesh Kumar also scored well, indicating their significant contributions. The scores reflect a combination of batting, bowling, and all-round performances throughout the season. Lower scores for some players suggest either limited opportunities or underperformance, while a few players received no score, due to lack of playing time or poor performance.

Batting Strength

1. Rishabh Pant led the batting charts with 446 runs at an impressive average of 40.55 and a strike rate of 155.4.

2. Tristan Stubbs showed excellent form, scoring 378 runs at a high average of 54 and a strike rate of 190.9.
3. Jake Fraser-McGurk emerged as an explosive batsman, scoring 330 runs at a strike rate of 234.04.
4. Abishek Porel contributed significantly with 327 runs at a strike rate of 159.51.

Bowling Strength:

1. Kuldeep Yadav was the standout bowler, taking 16 wickets at an average of 23.37 and an economy of 8.69.
2. Mukesh Kumar impressed with 17 wickets at an average of 21.64.
3. Axar Patel contributed with 11 wickets with a good economy of 7.65.
4. Khaleel Ahmed took 17 wickets with a decent economy of 9.58.

All-Round Performance:

1. Axar Patel showcased his all-round abilities, scoring 235 runs and took 11 wickets.
2. Tristan Stubbs, primarily a batsman, also took 3 wickets.

Emerging Talent:

1. Jake Fraser-McGurk stood out as a promising talent with his explosive batting.
2. Abishek Porel showed potential as a consistent run-scorer.
3. Rasikh Salam took 9 wickets in 8 matches.

The team's strength clearly lies in its batting, with multiple players capable of scoring quickly. The bowling unit, led by Kuldeep Yadav and supported by Mukesh Kumar and Axar Patel, also performed well. The emergence of young talents like Fraser-McGurk and Porel adds depth to the squad.

Key factors in DC's decision-making process include:

1. Rishabh Pant's leadership and batting prowess
2. The all-round abilities of Axar Patel
3. Kuldeep Yadav's consistent spin bowling performances
4. The explosive batting potential of Jake Fraser-McGurk
5. Tristan Stubbs' impressive batting in the previous season

7. Conclusion

7.1 Squad Optimization

Note: As of August 2024, there is no new information regarding player retention rules or the Right to Match (RTM) policy for IPL 2025. This analysis is based on the 2024 rules. Additionally, the model predictions used here will not impact future decisions or changes.

7.2 KKR Squad Optimization and picking best squad

7.2.1 Current players Overall score Prediction

Table 12 - KKR current players predicted overall score

Player	Predicted Overall score
Andre Russell	62.681908
Angkrish Raghuvanshi	43.662876
Anukul Roy	0.01888789
Harshit Rana	74.76428
Manish Pandey	18.542007
Mitchell Starc	66.87246
Nitish Rana	15.08972
Phil Salt	68.24854
Rahmanullah Gurbaz	18.410284
Ramandeep Singh	47.748146
Rinku Singh	40.41537
Shreyas Iyer	62.031254
Sunil Narine	65.13953
Vaibhav Arora	54.331146
Varun Chakaravarthy	74.76428
Venkatesh Iyer	62.110737

Based on the XGBoost model predictions and team dynamics, here's analysis of Kolkata Knight Riders' (KKR) potential retention strategy for IPL 2025: Core Retentions:

1. Sunil Narine (65.14)
2. Andre Russell (62.68)
3. Shreyas Iyer (62.03) - Captain
4. Varun Chakaravarthy (74.76)

KKR's retention strategy likely prioritizes a blend of consistent performers and recent standouts. Narine and Russell, with their high predicted scores and long-standing contributions to the franchise, are prime candidates. Shreyas Iyer, as the current captain and a solid middle-order batsman, provides leadership continuity. Varun Chakaravarthy's top predicted score and impressive bowling performances make him an asset for the team's bowling attack. Right to Match (RTM) Options:

1. Venkatesh Iyer (62.11)
2. Rinku Singh (40.42)

The RTM card could be used on Venkatesh Iyer, given his versatility and strong predicted performance. Rinku Singh, despite a lower predicted score, has shown potential as a finisher and could be a strategic RTM pick based on his past performances and future potential. Difficult Decisions:

- Phil Salt (68.25) and Mitchell Starc (66.87), despite their high predicted scores and contributions, may be released due to the limit on foreign player retentions.

- Nitish Rana (15.09), though injured in 2024 and having a low predicted score, might still be considered for RTM based on his past performances and experience with the team.

Potential Releases:

- Rahmanullah Gurbaz (18.41)
- Ramandeep Singh (47.75)
- Vaibhav Arora (54.33)
- Angkrish Raghuvanshi (43.66)

These players, while showing promise with their predicted scores, may not fit into the retention strategy given the limited slots available and the need to maintain a balanced squad. This approach balances maintaining the core team with strategic decisions for future success. The management faces tough choices, particularly regarding foreign players and emerging talents, as they aim to build a competitive squad for IPL 2025. The predicted scores provide valuable insight, but the final decisions will also consider factors such as team chemistry, player roles, and long-term strategy.

7.2.2 Potential Squad Options for KKR

The squad of KKR in 2024 contains 23 players with 8 overseas players, with 6 uncapped players. Including 9 batsman with 3 wicket keepers, 4 all-rounders and 10 bowlers. Out of these players 16 players have contributed for teams' success. Hence the squad is suggested based on the team possible retention, possible link to the team and available players and predicted data.

Note: The prediction is based on data is taken from 2021 to 2024

Suggested Squad Options for KKR IPL 2025

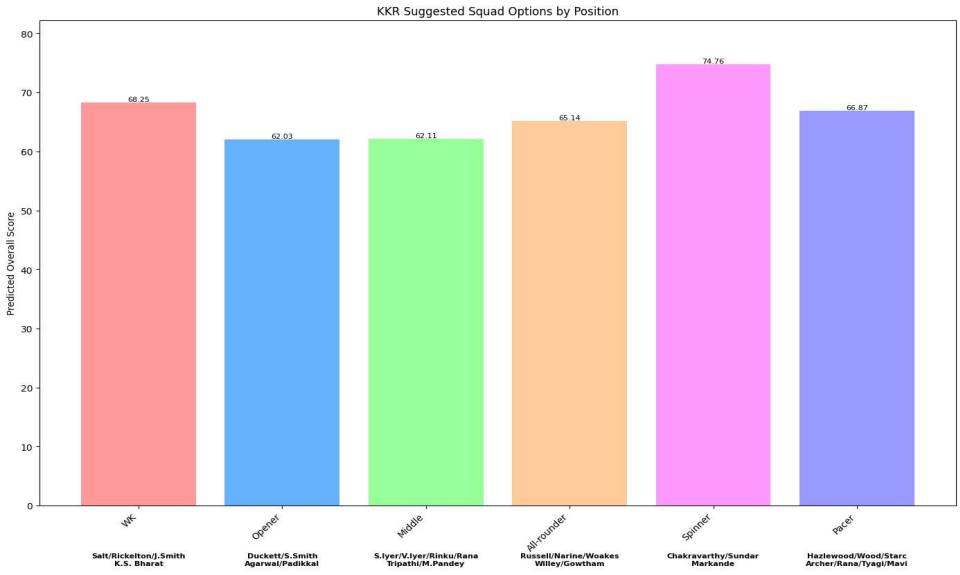


Figure 18 - Bar chart for squad options KKR

Overseas options include wicketkeepers Phil Salt and Ryan Rickelton, alongside batsmen Ben Duckett and Steve Smith. All-rounders Andre Russell, Sunil Narine, Chris Woakes, and David Willey offer versatility. The bowling attack features Josh Hazlewood, Mark Wood, Mitchell Starc, and Jofra Archer, with emerging talents like Atkinson and Potts.

Domestic choices highlight captain Shreyas Iyer, with K.S. Bharat as wicketkeeper. Batting strength comes from Venkatesh Iyer, Rinku Singh, Nitish Rana, Mayank Agarwal, Devdutt Padikkal, Rahul Tripathi, and Manish Pandey. All-rounders Washington Sundar, Krishappa Gowtham, and Shardul Thakur provide balance.

The bowling lineup includes promising pacers Harshit Rana, Karthik Tyagi, Shivam Mavi, and Mohsin Khan, alongside experienced options like Sandeep Warrier. Spin options feature Varun Chakravarthy and Mayank Markande ("see Appendix 5")

This suggested squad options for KKR in IPL 2025 are based on predicted performance data, potential retention strategies, and team dynamics. The focus is on creating a balanced and competitive team that leverages both international experience and domestic talent, ensuring KKR remains a formidable force in the league.

7.3 Delhi Capitals Squad Optimization and picking best squad

7.3.1 Current players Overall score Prediction

Table 13 - DC current players predicted overall scores

Player	Predicted Overall score
Mukesh Kumar	79.07398
Khaleel Ahmed	75.189224
Rishabh Pant	70.06949
Ishant Sharma	63.152454
Jake Fraser - McGurk	61.636425
Abishek Porel	60.6897
Tristan Stubbs	58.139206
Kuldeep Yadav	57.190998
Axar Patel	56.546555
Anrich Nortje	52.347652
Prithvi Shaw	45.95865
Shai Hope	44.87039
David Warner	43.445335
Rasikh Salam	40.407764
Mitchell Marsh	10.7273855
Gulbadin Naib	1.8736305
Ricky Bhui	0.2515321
Kumar Kushagra	0.096818216
Sumit Kumar	0.006852619
Jhye Richardson	-0.22722892
Lalit Yadav	-0.26193994
Lizaad Williams	-0.60568804

Based on the XGBoost model predictions and team dynamics, here's analysis of Delhi Capitals' (DC) potential retention strategy for IPL 2025:

Core Retentions:

- Rishabh Pant (70.07) - Captain and wicketkeeper-batsman

- Axar Patel (56.55) - All-rounder and consistent performer
- Jake Fraser-McGurk (61.64) - Explosive opener
- Kuldeep Yadav (57.19) - Key spinner

Right to Match (RTM) Options:

- Mukesh Kumar (79.07) - Highest predicted score
- Khaleel Ahmed (75.19) - Second-highest predicted score
- Tristan Stubbs (58.14) - Potential Player
- Abishek Porel (60.68) - Young talent

This revised strategy aligns better with the search results and acknowledges Stubbs' potential. The inclusion of Stubbs in the RTM list allows DC to potentially retain a player who has shown exceptional finishing skills and could be a long-term asset.

Difficult Decisions:

- Ishant Sharma (63.15), Anrich Nortje (52.35), and Prithvi Shaw (45.96) might still be released to create room for new strategies.

Potential Releases:

- David Warner (43.45)
- Shai Hope (44.87)
- Mitchell Marsh (10.73)

This approach balances retaining key performers, securing young talent with high potential, and creating opportunities for significant changes in the squad. It addresses DC's need to move from an average team to a top contender by making strategic decisions that combine experience (Pant, Axar) with emerging talents (Fraser-McGurk, Stubbs, Porel).

7.3.2 Potential Squad Formation for Delhi Capitals

The squad of DC in 2024 contains 27 players with 8 overseas players, with 11 uncapped players. Including 12 batsman with 6 wicket keepers, 5 all-rounders and 10 bowlers. Out of these players 22 players have contributed for team and out of 22, 7 players performed very poor. Hence the squad is suggested based on the team possible retention, possible link to the team and available players and predicted data.

Note: The suggestion is based on data is taken from 2021 to 2024

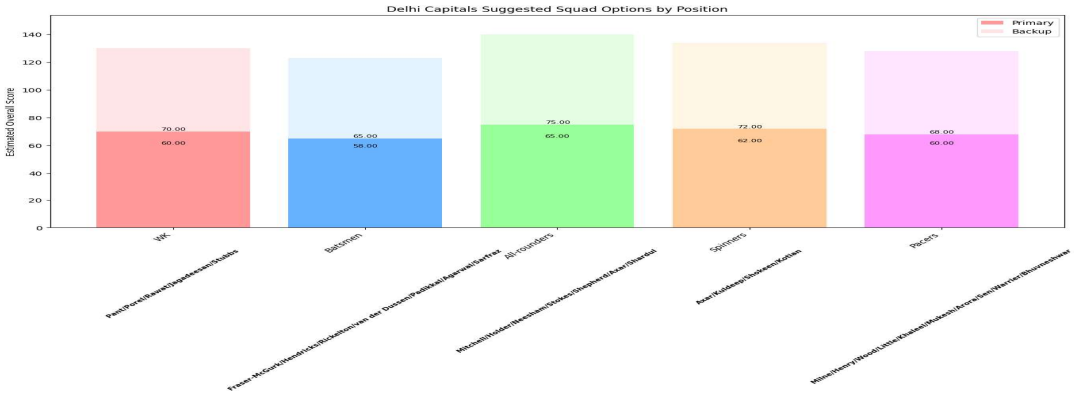


Figure 19 - Bar chart for squad options DC

Overseas options feature Jake Fraser-McGurk, an explosive batsman; Reeza Hendricks, a consistent T20 performer; and Tristan Stubbs, a dynamic middle-order batsman. Rassie van der Dussen brings experience, while all-rounders like Daryl Mitchell, Jason Holder, Jimmy Neesham, Ben Stokes, and Romario Shepherd add versatility. Fast bowlers include Adam Milne, Matt Henry, Mark Wood, and emerging talent Joshua Little.

Domestic choices highlight captain Rishabh Pant alongside wicketkeepers Abishek Porel, Anuj Rawat, and N. Jagadeesan. Batting strength comes from Devdutt Padikkal, Mayank Agarwal, and domestic star Sarfraz Khan. All-rounders like Axar Patel and Shardul Thakur provide balance.

The bowling attack features experienced pacer Bhuvneshwar Kumar, along with left-arm pacer Khaleel Ahmed and emerging talents like Mukesh Kumar, Vaibhav Arora, and Kuldeep Sen. Spin options include Kuldeep Yadav and emerging spinner Hrithik Shokeen (“see Appendix 6”).

8. Findings and Insights of Players and their performance scores

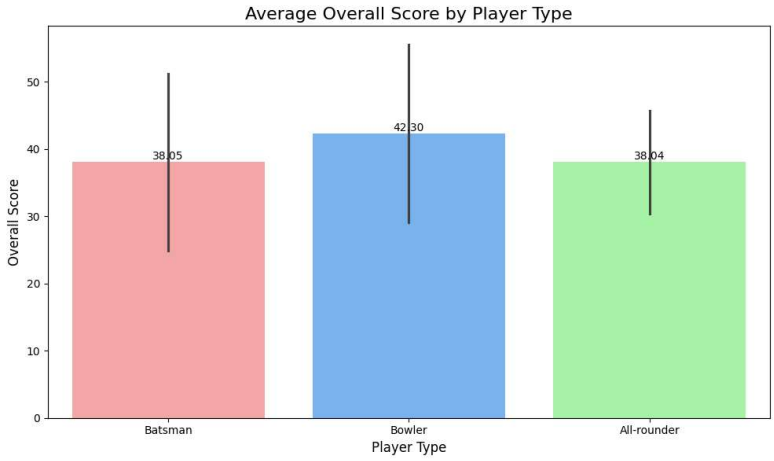


Figure 20 - Average overall score chart

This bar chart displays that bowlers have the highest average overall score, significantly higher than both batsmen and all-rounders. Interestingly, batsmen and all-rounders have very similar average scores, with batsmen only slightly outperforming all-rounders by 0.01 points.

8.1 Distribution of Overall Scores by Player Type

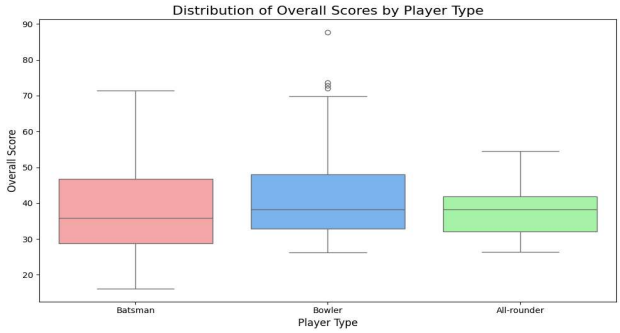


Figure 21 - Distribution of overall score by player type

1. The distribution shows that bowlers tend to perform better in terms of Overall score compared to the other two player types.

2. The similarity between batsmen and all-rounders' scores suggests that all-rounders are not necessarily at a disadvantage in terms of overall performance despite having to excel in both batting and bowling.

3. The highest individual Overall_score mentioned in the data is for YS Chahal, a bowler, with 87.70, which aligns with the higher average for bowlers.

8.2 Players with more than 300 runs with strike rate more than 130

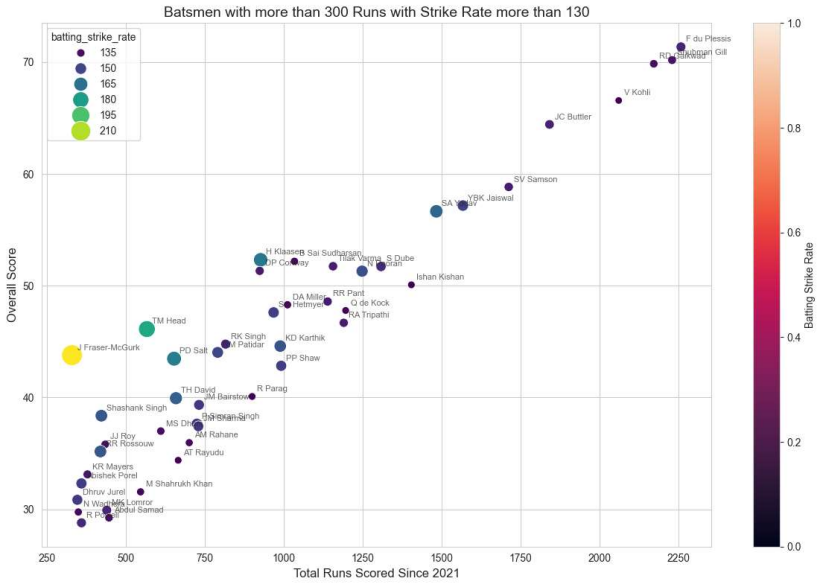


Figure 22 - Scatter plot for Batsman

This scatter plot illustrates the analysis of high-performing batsmen reveals notable players with impressive statistics. Jos Buttler an extraordinary strike rate of 158.62, showcasing his explosive batting style. Other key players include F du Plessis with 2257 runs at a strike rate of 141.59, and Shubman Gill, who has 2229 runs with a strike rate of 136.83. Additionally, T Head, Abhishek Sharma, and H Klassen also demonstrate strong performances, with strike rates exceeding 140. Fraser and Salt contribute to the aggressive batting lineup, emphasizing the importance of scoring quickly. These players exemplify a combination of high run totals and aggressive strike rates, making them valuable assets in competitive cricket, capable of changing the course of a match with their batting prowess.

8.3 Top All-rounders analysis

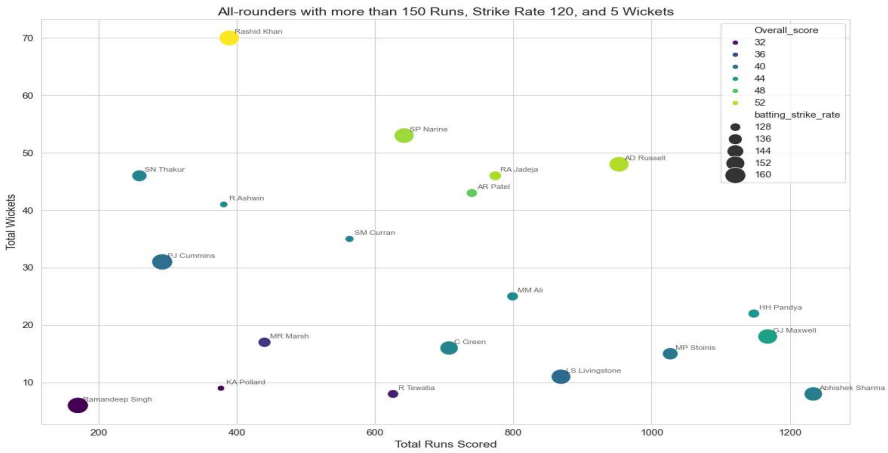


Figure 23 - Scatter plot for top all-rounders

This plot reveals several high-performing all-rounders in T20 cricket. Players like Rashid Khan, Andre Russell, Sunil Narine and Ravindra Jadeja stand out with impressive overall scores above 50. These players excel in both batting and bowling aspects of the game.

Rashid Khan leads the pack with an overall score of 54.49, showcasing his exceptional bowling skills combined with useful batting contributions. Andre Russell and Ravindra Jadeja follow closely, known for their explosive batting and crucial wicket-taking abilities.

Other notable performers include Harshal Patel, and Axar Patel, all scoring above 48. These players demonstrate the valuable combination of aggressive batting (high strike rates) and effective bowling (wicket-taking ability and economy).

The scatter plot effectively visualises the balance between run-scoring and wicket-taking abilities of these all-rounders, with the added dimension of strike rate represented by point size.

8.4 Top Economical Bowlers Analysis

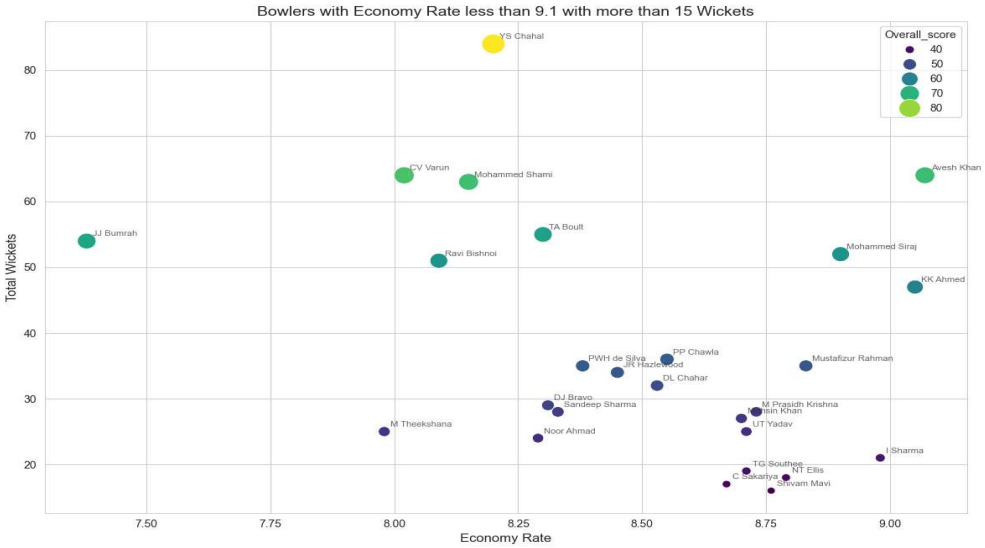


Figure 24 - Scatter plot for top economical bowlers

The analysis of top bowlers in T20 cricket reveals a group of exceptional performers who combine wicket-taking prowess with economical bowling. YS Chahal leads the pack with an impressive 84 wickets and an overall score of 87.70, showcasing his dominance in the format. The list features a mix of spin and pace bowlers, including standouts like CV Varun, Mohammed Shami, and Jasprit Bumrah. Notably, Bumrah boasts the best economy rate at 7.38.

8.5 Density distribution of overall scores:

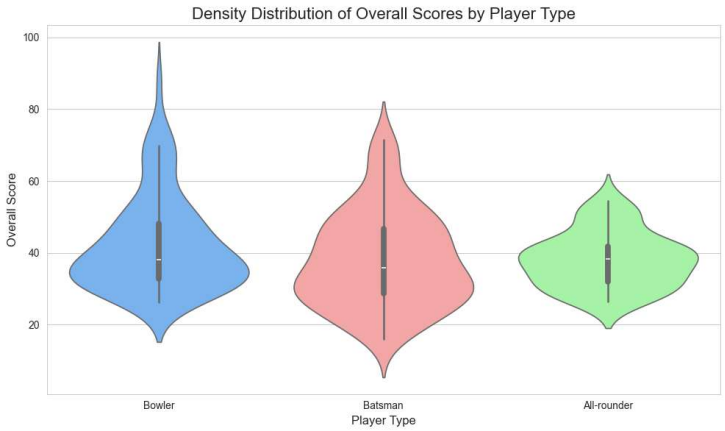


Figure 25 - Density distribution by player types.

The violin plot illustrates overall score distributions by player type. Batsmen likely show a wider spread with higher median scores, reflecting diverse roles and run-scoring focus. Bowlers may display a more compact distribution with lower median scores, indicating consistent, specialized performances. All-rounders potentially exhibit a broad range with median scores between batsmen and bowlers, representing their dual contributions. This visualization effectively captures the distinct performance characteristics of each player type in cricket.

8.6 Performance metrics of All-rounders

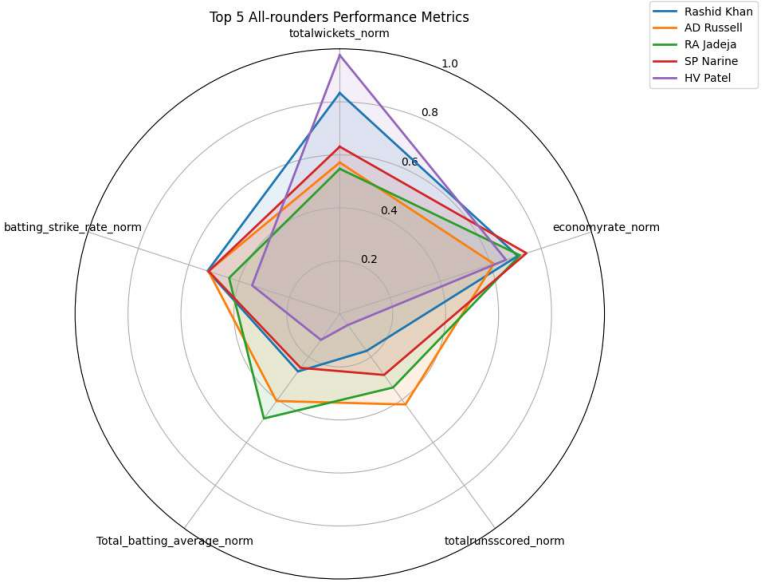


Figure 26 - Performance metrics of top 5 all-rounders

The radar chart effectively compares top all-rounders' strengths and weaknesses. Rashid Khan excels in bowling with high wicket-taking ability and good economy. Andre Russell shines as an aggressive batsman with useful bowling skills. Ravindra Jadeja offers a balanced performance with strong bowling economy and consistent batting. Sunil Narine is primarily a bowler with excellent economy and the ability to score quick runs. Harshal Patel stands out as a bowling all-rounder with strong wicket-taking ability and moderate batting contributions. This visualization provides an intuitive understanding of each player's performance profile across multiple cricket aspects.

8.7 Research Conclusion

The conclusion of the research project emphasizes the significance of strategic squad optimization for the Kolkata Knight Riders (KKR) and Delhi Capitals (DC) in the context of the upcoming IPL mega auction in 2025.

Key findings indicate that KKR's successful championship strategies in 2024 stemmed from effective player retention and utilization, while DC's potential remains underutilized despite having a strong young core. The study utilized various machine learning models to analyse player performance and predict outcomes, revealing critical insights into team dynamics and performance metrics.

The research highlights the importance of quantitative analysis in sports, offering a framework for teams to enhance decision-making processes regarding player selection and strategic planning. By focusing on the unique challenges faced by both teams, the study provides actionable recommendations for optimizing squad composition, particularly for DC in leveraging their young talent effectively.

Overall, this research contributes to the broader field of quantitative sports analytics, offering valuable insights not only for KKR and DC but also for other T20 franchises globally. It underscores the evolving nature of team management in cricket, advocating for data-driven approaches to improve performance and competitiveness in the IPL.

9. Recommendations

Based on the comprehensive research on optimizing squad compositions for IPL teams, particularly focusing on Kolkata Knight Riders (KKR) and Delhi Capitals (DC), here are some key recommendations:

1. Embrace data-driven decision making: The IPL is evolving rapidly. Teams should leverage the power of analytics to make smarter choices in player selection and strategy formulation. This approach can provide valuable insights that might not be apparent to the naked eye.

2. Nurture young talent strategically: While it's tempting to always go for established stars, don't underestimate the potential of young players. Develop a system to identify and groom emerging talents, giving them the right opportunities to shine. This is especially crucial for teams like DC, which has a wealth of young talent waiting to be unleashed.

3. Balance squad wisely: Cricket is a game of balance, and so is team composition. Aim for a mix of experienced veterans and energetic youngsters, aggressive hitters and steady anchors, pace bowlers and crafty spinners. This diversity can help teams adapt to various match situations and conditions.

4. Invest in multi-dimensional players: In T20 cricket. Players who can contribute to multiple areas – be it batting, bowling, or fielding – can be game-changers. They provide captains with more options and can turn matches on their head.

5. Stay adaptable: The IPL is a long tournament with changing conditions. Teams that can quickly adapt their strategies based on performance data and match situations often come out on top. Flexibility in approach can be a key differentiator.

6. Optimize the auction strategy: With the mega auction coming up, use predictive models to inform bidding decisions. Focus on players who not only have good historical stats but also show potential for growth and fit well within the team's overall strategy.

7. Foster a culture of continuous improvement: Encourage players and coaching staff to regularly review performance data and work on areas of improvement. Create an environment where everyone is committed to getting better every day.

8. Look beyond the boundaries: While focusing on the IPL, keep an eye on performances in other T20 leagues worldwide. This global perspective can help in identifying undervalued players who might become match-winners for a team.

While data and analytics are powerful tools, cricket is still a human game. The most successful teams will be those that can blend analytical insights with the intangibles of team spirit, leadership, and on-field chemistry. By implementing these recommendations, teams can position themselves for success in the highly competitive world of the Cricket.

10. References

1. Amala Kaviya, V.S., Mishra, A.S. and Valarmathi, B. (2020) 'Comprehensive Data Analysis and Prediction on IPL using Machine Learning Algorithms', International Journal on Emerging Technologies, 11(3), pp. 218-228. (Accessed: 15 August 2024).
2. Bajaj, A. (2023) 'Prediction of Player Performance for IPL and Analyzing the Attributes Involved, Using Explainable AI', MSc Research Project, National College of Ireland. Available at: https://norma.ncirl.ie/6564/1/ayushibajaj.pdf (Accessed: 15 August 2024).
3. Berrar, D., Lopes, P. and Dubitzky, W. (2019). Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine Learning, 108(1), pp.97-126.
4. Board of Control for Cricket in India (2023) Indian Premier League. Available at: https://www.iplt20.com/ (Accessed: 15 August 2024).
5. Brownlee, J., 2018. How to Use ROC Curves and Precision-Recall Curves for Classification in Python. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/ [Accessed 25 August 2024].
6. Bunker, R.P. and Thabtah, F., 2019. A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), pp.27-33.
7. Caya, O. and Bourdon, A., 2016. A framework of value creation from business intelligence and analytics in competitive sports. In 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 1061-1071). IEEE.
8. Cervone, D., D'Amour, A., Bornn, L. and Goldsberry, K., 2016. A multiresolution stochastic process model for predicting basketball possession outcomes. Journal of the American Statistical Association, 111(514), pp.585-599.

9. Cherkassky, V. and Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks, 17(1), pp.113-126.
10. Colwell, D., Jones, B. and Gillett, J. (1991) “75.7 A Markov Chain in Cricket (MCC!),” The Mathematical Gazette, 75(472), pp. 183–185. Available at: https://doi.org/10.2307/3620249.
11. Duff & Phelps (2022) IPL Brand Valuation Report 2022. Mumbai: Duff & Phelps.
12. Drucker, H., Burges, C.J., Kaufman, L., Smola, A.J. and Vapnik, V., 1997. Support vector regression machines. Advances in Neural Information Processing Systems, 9, pp.155-161.
13. Economic Times (2023) 'IPL becomes decacorn, valuation soars 75% since 2020', 27 December.
14. ESPN Cricinfo (2023) Indian Premier League. Available at: https://www.espncricinfo.com/series/indian-premier-league-2023-1345038 (Accessed: 15 August 2024).
15. ESPNcricinfo (2024). How KKR shaped themselves into the awesome class of 2024. [online] Available at: https://www.espncricinfo.com/story/ipl-2024-final-kkr-vs-srh-how-kkr-shaped-themselves-into-the-awesome-class-of-2024-1435320 [Accessed 15 Aug. 2024].
16. Fried, G. and Mumcu, C. eds., 2016. Sport analytics: A data-driven approach to sport business and management. Taylor & Francis.
17. Hajian-Tilaki, K., 2013. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian Journal of Internal Medicine, 4(2), pp.627-635.
18. Hubáček, O., Šourek, G. and Železný, F. (2019). Exploiting sports-betting market using machine learning. International Journal of Forecasting, 35(2), pp.783-796.
19. Ishi, M., Patil, D.J., Patil, D.N. and Patil, D.V. (2022) 'Winner Prediction in One Day International Cricket Matches Using Machine Learning Framework: An Ensemble Approach', Indian Journal of Computer Science and Engineering, 13, pp. 628–641.
20. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning. New York: Springer.
21. JioCinema (2023) 'IPL 2023 Final Sets Global Streaming Record', Press Release, 30 May.
22. Kadapa, S. (2013) 'How Sustainable is the Strategy of the Indian Premier League-IPL? A Critical Review of 10 Key Issues That Impact the IPL Strategy', International Journal of Scientific and Research Publications, 3.
23. Kemper, C. and Breuer, C., 2016. How efficient is dynamic pricing for sport events? Designing a dynamic pricing model for Bayern Munich. International Journal of Sport Finance, 11(1), pp.4-25.
24. Liu, G., Luo, Y., Schulte, O. and Kharrat, T., 2020. Deep soccer analytics: learning an action-value function for evaluating soccer players. Data Mining and Knowledge Discovery, 34(5), pp.1531-1559.
25. Loland, S., 2018. Performance-enhancing drugs, sport, and the ideal of natural athletic performance. The American Journal of Bioethics, 18(6), pp.8-15.
26. McHale, I.G., Scarf, P.A. and Folker, D.E., 2012. On the development of a soccer player performance rating system for the English Premier League. Interfaces, 42(4), pp.339-351.
27. Memmert, D. and Raabe, D., 2018. Data analytics in football: Positional data collection, modelling and analysis. Routledge.
28. Ofoghi, B., Zeleznikow, J., MacMahon, C. and Raab, M., 2013. Data mining in elite sports: a review and a framework. Measurement in Physical Education and Exercise Science, 17(3), pp.171-186.
29. Peacock, R.H. (1950) “2124. The New Ball in Cricket,” The Mathematical Gazette, 34(307), pp. 58–60. Available at: https://doi.org/10.2307/3610894.

30. Prakash, A., Ghosh, A. and Guha, B. (2019) 'Player Ranking System for IPL Using Machine Learning', International Journal of Sports Analytics, 5(1), pp. 1-12.
31. Rodrigues, M., Vinay, S., Naik, N., Deshpande, S. and Samant, S. (2019). Data visualization and toss related analysis of IPL teams and batsmen performances. [online] ResearchGate.
32. Rommers, N., Rössler, R., Goossens, L., Vaeyens, R., Lenoir, M., Witvrouw, E. and D'Hondt, E., 2020. Risk of acute and overuse injuries in youth elite soccer players: Body size and growth matter. Journal of Science and Medicine in Sport, 23(3), pp.246-251.
33. Rossi, A., Pappalardo, L., Cintia, P., Iaia, F.M., Fernàndez, J. and Medina, D., 2018. Effective injury forecasting in soccer with GPS training data and machine learning. PloS one, 13(7), p.e0201264.
34. Shah, J. (2023) The IPL Story: Cricket, Commerce and Glamour. New Delhi: Rupa Publications.
35. Shah, R., Ghosh, A. and Guha, B. (2016) 'IPL 2016: A Comprehensive Analysis of the Performance of Teams', International Journal of Sports Analytics, 2(1), pp. 1-15.
36. Seshadri, D.R., Drummond, C., Craker, J., Rowbottom, J.R. and Voos, J.E., 2019. Wearable devices for sports: New integrated technologies allow coaches, physicians, and trainers to better understand the physical demands of athletes in real time. IEEE pulse, 10(1), pp.38-43.
37. Smola, A.J. and Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing, 14(3), pp.199-222.
38. Sportstar (2023) 'IPL media rights sold for Rs 48,390 crore: Disney Star retains TV rights, Viacom18 bags digital package', The Hindu, 14 June.
39. Thomas, G., Gade, R., Moeslund, T.B., Carr, P. and Hilton, A., 2017. Computer vision for sports: Current applications and research topics. Computer Vision and Image Understanding, 159, pp.3-18.
40. IPL Governing Council (2024) IPL 2024: Playing Conditions. Mumbai: BCCI.
41. Delhi Capitals (2023) Official Website. Available at: https://www.delhicapitals.in/ (Accessed: 15 August 2024).
42. Gujarat Titans (2023) Official Website. Available at: https://www.gujarattitansipl.com/ (Accessed: 15 August 2024).
43. Kolkata Knight Riders (2023) Official Website. Available at: https://www.kkr.in/ (Accessed: 15 August 2024).
44. Lucknow Super Giants (2023) Official Website. Available at: https://www.lucknowsupergiants.in/ (Accessed: 15 August 2024).
45. Mumbai Indians (2023) Official Website. Available at: https://www.mumbaiindians.com/ (Accessed: 15 August 2024).
46. Punjab Kings (2023) Official Website. Available at: https://www.punjabkingsipl.in/ (Accessed: 15 August 2024).
47. Rajasthan Royals (2023) Official Website. Available at: https://www.rajasthanroyals.com/ (Accessed: 15 August 2024).
48. Royal Challengers Bangalore (2023) Official Website. Available at: https://www.royalchallengers.com/ (Accessed: 15 August 2024).
49. Sunrisers Hyderabad (2023) Official Website. Available at: https://www.sunrisershyderabad.in/ (Accessed: 15 August 2024).

Appendices

Appendix 1 – About IPL Teams

The Indian Premier League (IPL) currently features ten franchise teams, each representing different cities or states across India (Board of Control for Cricket in India, 2023):

Figure 277 - Chennai Super Kings Logo

Chennai Super Kings (CSK): Known for their consistency and led by the iconic MS Dhoni, CSK has won four IPL titles (ESPN Cricinfo, 2023).

Figure 28 - Delhi Capitals Logo

Delhi Capitals (DC): Formerly Delhi Daredevils, this team rebranded in 2018 and has been building a strong young core of Indian talent (Delhi Capitals, 2023).

Figure 29 - Gujarat Titans Logo

Gujarat Titans (GT): One of the newest additions to the IPL, joining in 2022, they made an immediate impact by winning the title in their debut season (Gujarat Titans, 2023).

Figure 30 - Kolkata Knight Riders Logo

Kolkata Knight Riders (KKR): Co-owned by Bollywood star Shah Rukh Khan, KKR has won two IPL titles and has a massive fan following (Kolkata Knight Riders, 2023).

Figure 31 - Lucknow Super Giants Logo

Lucknow Super Giants (LSG): Another new franchise that joined in 2022, they've quickly established themselves as strong contenders (Lucknow Super Giants, 2023).

Figure 32 - Mumbai Indians Logo

Mumbai Indians (MI): The most successful IPL team with five titles, MI is known for its star-studded lineup and ability to nurture young talent (Mumbai Indians, 2023).

Figure 33 - Punjab Kings Logo

Punjab Kings (PBKS): Formerly Kings XI Punjab, this team rebranded in 2021 and is still seeking its first IPL title (Punjab Kings, 2023).

Figure 34 - Rajasthan Royals Logo

Rajasthan Royals (RR): The inaugural IPL champions in 2008, RR is known for its ability to unearth and develop lesser-known players (Rajasthan Royals, 2023).

Figure 35 - Royal Challengers Bengaluru logo

Royal Challengers Bangalore (RCB): Despite boasting some of cricket's biggest names, RCB is still chasing their first IPL title (Royal Challengers Bangalore, 2023).

Figure 36 - Sunrisers Hyderabad

Sunrisers Hyderabad (SRH): Known for their strong bowling attacks, SRH won the title in 2016 and has consistently been a playoff contender (Sunrisers Hyderabad, 2023).

Appendix 2 – Team Performance

Mumbai Indians (MI):

Mumbai Indians have played the most matches (261) and won the most games (144) in IPL. Their success is evident in their 5 IPL titles, the highest among all teams. They've reached the finals 6 times and made it to the playoffs 11 times, showcasing their consistency (Board of Control for Cricket in India, 2023).

Royal Challengers Bangalore (RCB):

Despite playing 256 matches and winning 123, RCB has never won an IPL title. They've reached the finals 3 times and made the playoffs 9 times. Their inability to convert playoff appearances into titles has been a point of discussion among cricket analysts (ESPN Cricinfo, 2023).

Kolkata Knight Riders (KKR):

KKR has played 252 matches, winning 131. They've clinched 3 IPL titles and reached the finals 4 times. With 7 playoff appearances, they've shown consistency in reaching the later stages of the tournament (Kolkata Knight Riders, 2023).

Delhi Capitals (DC):

Formerly Delhi Daredevils, DC has played 252 matches but won only 115. They've never won an IPL title and have reached the finals only once. With 6 playoff appearances, they've struggled to make a significant impact in the tournament's history (Delhi Capitals, 2023).

Punjab Kings (PK):

PK has played 246 matches, winning 112. They've never won an IPL title and have reached the finals only once. With just 2 playoff appearances, they've been one of the less successful teams in the IPL (Punjab Kings, 2023).

Chennai Super Kings (CSK):

Despite playing fewer matches (239) than some other teams, CSK has been incredibly successful. They've won 138 matches and 5 IPL titles, equaling MI's record. With 10 final appearances and 13 playoff qualifications, they're considered one of the most consistent teams in IPL history (Chennai Super Kings, 2023).

Rajasthan Royals (RR):

RR has played 222 matches, winning 112. They won the inaugural IPL in 2008 but haven't replicated that success since. With 2 final appearances and 5 playoff qualifications, they've had mixed fortunes in the tournament (Rajasthan Royals, 2023).

Sunrisers Hyderabad (SRH):

SRH entered the IPL later than the original teams but has made a significant impact. They've played 182 matches, winning 88. They've won 1 IPL title and reached the finals 3 times, with 6 playoff appearances (Sunrisers Hyderabad, 2023).

Gujarat Titans (GT):

As one of the newest teams, GT has played only 45 matches but has already won 28 of them. They won the IPL in their debut season in 2022 and reached the finals again in 2023, showing immediate success (Gujarat Titans, 2023).

Lucknow Super Giants (LSG):

Another new entrant, LSG, has played 44 matches and won 24. While they haven't reached a final yet, they've made it to the playoffs in both their seasons, indicating a strong start to their IPL journey (Lucknow Super Giants, 2023).

Appendix 3 – Reason for using Linear Regression

1. **Continuous Dependent Variable:**
   Dependent variable, Win_Ratio, is a continuous variable, which is suitable for linear regression analysis.
2. **Multiple Independent Variables:**
   The dataset includes multiple potential predictors (e.g., Played, Won, Lost, N/R, lost_Ratio, Titles, Finalists, Playoff), making multiple linear regression an appropriate choice.
3. **Relationship Exploration:**
   Linear regression can help identify which factors have the strongest influence on a team's win ratio, providing valuable insights into team performance.
4. **Performance Prediction:**
   The model can be used to predict a team's expected win ratio based on other performance metrics, which could be useful for team management and strategy planning.
5. **Quantifiable Impact:**
   Linear regression provides coefficients that quantify the impact of each independent variable on the win ratio, allowing for a clear understanding of each factor's importance.

6. **Model Interpretability:**
   In sports analytics, it's often crucial to have models that can be easily interpreted by coaches, managers, and other stakeholders. Linear regression provides this interpretability.

7. **Baseline Model:**
   Even if more complex models might be explored later, linear regression serves as an excellent baseline model to compare against more sophisticated approaches.

8. **Assumption Testing:**
   The dataset allows for testing various assumptions of linear regression (like linearity, homoscedasticity, and multicollinearity), which can provide insights into the data's structure.

9. **Small Dataset Handling:**
   With a relatively small dataset (10 observations), linear regression can still provide reliable results, whereas more complex models might overfit.

10. **Performance Metrics:**
    The high R-squared value (0.9969) suggests that linear regression is capturing a significant amount of variance in the win ratio, indicating a good fit for this data.

Appendix 4 – Reason for using Rule Based Scoring System

The rule-based scoring system for cricket aim to quantify a player's overall performance by assigning points based on various aspects of their game.

1. **Holistic Player Assessment:**
   - The code combines multiple performance metrics (runs, average, strike rate, wickets, economy rate) to create a comprehensive evaluation of each player.
   - This approach provides a more complete picture of a player's contribution than individual statistics alone.

2. **Role-Based Evaluation:**
   - By categorizing players as Batsmen, Bowlers, or All-rounders, the analysis acknowledges the different roles within a cricket team.
   - This allows for fair comparisons between players with similar roles and responsibilities.

3. **Normalised Comparisons:**
   - Normalising metrics enables fair comparisons across different scales (e.g., comparing runs scored with wickets taken).
   - This is essential for creating a unified scoring system that can be applied across diverse player types.

4. **Weighted Performance Metrics:**
   - Assigning weights to different metrics (e.g., giving more importance to total runs for batsmen or wickets for bowlers) reflects the relative importance of various aspects of performance.
   - This nuanced approach aligns the analysis with the strategic priorities of T20 cricket.

5. **Identifying All-Round Talent:**
    - The system's ability to evaluate all-rounders separately recognizes the unique value of players who contribute significantly in both batting and bowling.
6. **Ranking Within Categories:**
    - Ranking players within their specific roles (batsman, bowler, all-rounder) provides context-specific performance assessments.
    - This is valuable for team selection, strategy formulation, and player development.
7. **Data-Driven Decision Making:**
    - The analysis provides an objective, data-driven basis for decisions related to team composition, player retention, and strategic planning.
8. **Performance Benchmarking:**
    - By creating a standardized scoring system, teams can benchmark player performances across seasons or compare players from different teams.
9. **Talent Identification:**
    - This system can help identify undervalued players or rising talents who might not stand out in traditional statistics but perform well in this comprehensive analysis.
10. **Contract and Auction Strategies:**
    - For leagues like the IPL, this analysis can inform bidding strategies during player auctions and help in determining player values for contracts.
11. **Fan Engagement and Fantasy Sports:**
    - Providing a single, comprehensive score for each player enhances fan engagement and can be particularly useful for fantasy cricket leagues.
12. **Continuous Performance Monitoring:**
    - This type of analysis can be easily updated with new match data, allowing for continuous monitoring of player performance throughout a season or across multiple seasons.

Appendix 4 – Dataset variables

1. **match_id**: This column contains a unique identifier for each match, allowing for easy referencing and data management. It helps distinguish between different matches in the dataset.
2. **season**: This indicates the specific IPL season during which the match took place. It typically refers to the year of the tournament, providing context for the data.
3. **start_date**: This column records the date on which the match commenced. It is essential for temporal analysis, allowing researchers to study trends over different seasons or specific time periods.

4. **venue**: This specifies the location where the match was held. Knowing the venue is important for analysing home advantage, pitch conditions, and crowd influence on the game.
5. **innings**: This indicates whether the data pertains to the first or second innings of the match. In cricket, each team bats for one or two innings, and this column helps differentiate between them.
6. **ball**: This column records the specific ball number within the over. It provides granular detail about the match, allowing for in-depth analysis of individual deliveries.
7. **batting_team**: This specifies the team that is currently batting during the delivery. It is crucial for understanding team performance and strategies.
8. **bowling_team**: This indicates the team that is currently bowling. This information is essential for analysing bowling strategies and effectiveness.
9. **striker**: This column names the batsman facing the current delivery. It is important for analysing individual player performance and contributions.
10. **non_striker**: This indicates the batsman at the other end of the pitch who is not facing the current delivery. It provides context for partnerships and running between the wickets.
11. **extras**: This column records the total extra runs scored on that delivery, which can include wides, no-balls, and other extras. It is important for assessing the impact of extras on the match outcome.
12. **wides**: This specifies the number of wide balls bowled during that delivery. Wides contribute to the extras and can affect the match's flow and scoring.
13. **noballs**: This indicates the number of no-balls bowled on that delivery. No-balls also contribute to extras and can lead to free hits, impacting scoring opportunities.
14. **byes**: This column records the number of byes scored on that delivery, which occur when the ball passes the wicketkeeper without touching the bat or body of the batsman.
15. **legbyes**: This specifies the number of leg byes scored, which occur when the ball hits the batsman's body (excluding the hand) and runs are taken.
16. **penalty**: This column records any penalty runs awarded to the batting or bowling team, which can occur due to infractions by the fielding team.
17. **wicket_type**: This indicates the type of dismissal if a wicket fell on that delivery (e.g., bowled, caught, LBW). It is crucial for analysing how wickets are taken.
18. **player_dismissed**: This column names the player who was dismissed on that delivery, providing insight into key moments in the match.
19. **other_wicket_type**: This specifies any secondary wicket type, if applicable, for cases where multiple dismissals occur in a single delivery (e.g., run out).
20. **other_player_dismissed**: This column names any other player who was dismissed on that delivery, providing additional context for significant events.

Aston University
BIRMINGHAM UK

Appendix 5 – Potential squad Options for KKR

Overseas Players Options

- **Wicketkeepers**:
    - **Phil Salt**: An explosive batsman who can change the game.
    - **Rickelton**: A solid option with potential.
    - **Jamie Smith**: A young talent for future growth.
- **Batsmen**:
    - **Ben Duckett**: A dynamic player with a strong T20 record.
    - **Steve Smith**: An experienced batsman known for his technique and leadership.
- **All-rounders**:
    - **Andre Russell**: A key all-rounder with match-winning capabilities.
    - **Sunil Narine**: A long-time KKR asset with both batting and bowling skills.
    - **Chris Woakes**: Adds versatility and experience to the squad.
    - **David Willey**: Offers depth and balance as an all-rounder.
- **Bowlers**:
    - **Josh Hazlewood**: Known for his precision and effectiveness.
    - **Mark Wood**: Brings express pace and aggression.
    - **Mitchell Starc**: A premier fast bowler with the ability to take wickets.
    - **Atkinson**: A developing talent with potential.
    - **Potts**: An emerging bowler to consider.
    - **Jofra Archer**: A high-impact player with a proven track record.

Domestic Players Options

- **Wicketkeepers**:
    - **K.S. Bharat**: A reliable option for the wicketkeeping role.
- **Batsmen**:
    - **Shreyas Iyer**: The captain and a crucial middle-order batsman.
    - **Venkatesh Iyer**: Offers flexibility in the batting lineup.
    - **Rinku Singh**: A promising finisher with a bright future.
    - **Nitish Rana**: Experienced and capable of anchoring the innings.
    - **Mayank Agarwal**: Adds stability and experience.
    - **Devdutt Padikkal**: A young talent with strong potential.

- **Rahul Tripathi**: Known for his aggressive batting style.
- **Manish Pandey**: Brings experience and depth to the batting order.

- **All-rounders**:
  - **Washington Sundar**: Valuable for his bowling and batting skills.
  - **Krishappa Gowtham**: Adds depth and versatility.
  - **Shardul Thakur**: Known for his ability to contribute in multiple ways.
- **Bowlers**:
  - **Harshit Rana**: An emerging fast bowler with promise.
  - **Varun Chakravarthy**: A key spinner with wicket-taking ability.
  - **Karthik Tyagi**: Young and talented fast bowler.
  - **Shivam Mavi**: Known for his pace and skill.
  - **Sakariya**: Adds depth to the pace attack.
  - **Mohsin Khan**: A promising young bowler.
  - **Sandeep Warrier**: Experienced and reliable.
  - **Mayank Markande**: Spin option with experience.

Appendix 6 – Potential squad Options for DC

Overseas Players Options

- **Jake Fraser-McGurk**: A young talent with explosive batting capabilities, adding depth to the batting lineup.
- **Reeza Hendricks**: A consistent performer in T20 cricket, known for his ability to anchor innings and score quickly.
- **Ryan Rickelton**: An emerging batsman with a strong domestic record, capable of playing aggressive innings.
- **Tristan Stubbs**: A dynamic batsman with power-hitting skills, ideal for the middle order.
- **Rassie van der Dussen**: A seasoned international player known for his technique and ability to play under pressure.
- **Daryl Mitchell**: A versatile all-rounder who can contribute with both bat and ball, enhancing team balance.
- **Jason Holder**: An experienced all-rounder with a proven track record in T20s, offering both bowling and batting depth.
- **Jimmy Neesham**: A dynamic all-rounder known for his big-hitting ability and useful seam bowling.
- **Ben Stokes**: A match-winner with exceptional all-round skills, capable of changing games single-handedly.

- **Romario Shepherd**: A powerful all-rounder who can contribute significantly with the bat and provide pace bowling options.
- **Adam Milne**: A fast bowler with express pace, known for his wicket-taking ability in T20 cricket.
- **Matt Henry**: A skilled bowler with experience in international cricket, effective in both powerplays and death overs.
- **Mark Wood**: An aggressive fast bowler known for his pace and ability to take key wickets.
- **Joshua Little**: An emerging talent with potential as a left-arm fast bowler.

Domestic Players Options

- **Rishabh Pant**: The captain and wicketkeeper, known for his explosive batting and game-changing abilities.
- **Abishek Porel**: A promising wicketkeeper-batsman, providing depth in the lower order.
- **Anuj Rawat**: A young wicketkeeper with potential, looking to make an impact in the IPL.
- **N. Jagadeesan**: A reliable wicketkeeper-batsman with a solid domestic record.
- **Devdutt Padikkal**: A talented batsman with a strong ability to score quickly, adding firepower to the top order.
- **Mayank Agarwal**: An experienced opener known for his solid technique and ability to build innings.
- **Sarfraz Khan**: A domestic star with a strong record, capable of performing under pressure.
- **Axar Patel**: A key all-rounder known for his bowling and handy batting, providing balance to the team.
- **Shardul Thakur**: An all-rounder who can contribute with both bat and ball, known for his wicket-taking ability.
- **Khaleel Ahmed**: A left-arm pacer with experience in T20 cricket, effective in the powerplay.
- **Mukesh Kumar**: An emerging fast bowler with potential, looking to establish himself in the IPL.
- **Vaibhav Arora**: A promising young bowler with a good domestic record.
- **Kuldeep Sen**: A fast bowler with the ability to take wickets, adding depth to the bowling lineup.
- **Sandeep Warrier**: An experienced bowler providing additional options in the pace attack.
- **Bhuvneshwar Kumar**: A seasoned pacer known for his swing bowling and experience in high-pressure situations.
- **Tanush Kotian**: An all-rounder with potential, offering flexibility to the squad.
- **Kuldeep Yadav**: A skilled spinner known for his wicket-taking ability and variations.
- **Hrithik Shokeen**: An emerging spinner with potential to contribute to the middle overs.