Crickonomics: The Moneyball Strategy for IPL

A PROJECT REPORT SUBMITTED TO

SVKM'S NMIMS (DEEMED- TO- BE UNIVERSITY)

IN PARTIAL FULFILMENT FOR THE DEGREE OF

BACHELOR OF SCIENCE
IN
DATA SCIENCE

BY:

A001 SHREY AGARWAL

A002 SHASHVATH ARUN

A013 SHITIZ KUMAR GUTA

A022 KEEGAN NUNES

SVKM'S NMIMS Deemed-to-be UNIVERSITY

NILKAMAL SCHOOL OF MATHEMATICS,
APPLIED STATISTICS & ANALYTICS

NMIMS NSOMASA
Ground Floor, SBMP Phase I,
Irla, N. R. G Marg, Opposite Cooper Hospital,
Vile-Parle (West), Mumbai – 400 056

CERTIFICATE

This is to certify that the work described in this report entitled “Crickonomics: The Moneyball Strategy for IPL” has been carried out by Mr Shitiz Kumar Gupta, Mr. Shashvath Arun, Mr. Shrey Agarwal, and Mr. Keegan Nunes under my supervision. I certify that this is his/her bonafide work. The work described is original and has not been submitted for any degree to this or any other University.

Date: 07/11/2025
Place: Mumbai

Supervisor

(Prof. Pradnya Khandeparkar)

ACKNOWLEDGEMENT

First and foremost, we would like to express our heartfelt gratitude towards the Dean of NMIMS NSOMASA, Dr. Sushil Kulkarni, for your constant support and guidance throughout this project ‘Crickonomics: The Moneyball Strategy for IPL’.Your encouragement and feedback have been instrumental in helping us achieve our goals and complete this project successfully.

We would also like to thank our mentor, Prof. Pradnya Khandeparkar, for her valuable insights and assistance in navigating the complexities of this project. Her expertise in statistics has been a tremendous asset to us, and we have learned a lot under her guidance.

We would like to express our appreciation to other faculty members, our friends, and family members who have supported us throughout this project with their words of encouragement and motivation.

Once again, we extend our sincere thanks to Dean Sir and Pradnya Ma’am for providing us with this wonderful opportunity and for being an inspiration throughout our academic journeys.

Sincerely,
Keegan Nunes
Shitiz Gupta
Shashvath Arun
Shrey Agarwal

TABLE OF CONTENTS

4. ABSTRACT

This project applies the analytical principles popularized by the movie *Moneyball* to the context of the Indian Premier League (IPL), with the objective of developing a data-driven framework for team selection. In the competitive and financially intensive environment of the IPL, franchises often rely on reputation, intuition, and recent performances when selecting players. This study aims to demonstrate that a more systematic, evidence-based approach can identify undervalued talent and construct a balanced, cost-effective team capable of performing competitively.

Comprehensive player statistics were collected across batting, bowling, and fielding disciplines. Using this data, a structured player rating system was developed to evaluate players objectively based on their measured performances rather than subjective assessments. The framework integrates multiple statistical methods, including supervised machine learning techniques, feature engineering, and dimensionality reduction. Players were then rated according to their consistency, efficiency, and overall contribution to the game.

To evaluate the effectiveness of the selection strategy, the final team composed primarily of unsold or underrated players was assessed through Monte Carlo simulations, which modeled numerous match scenarios to estimate performance under varied conditions. The findings reveal that statistically optimized selection can produce teams with strong performance potential while maintaining financial efficiency. The project highlights the power of quantitative methods in sports management and suggests that the *Moneyball* philosophy, when adapted to cricket, can bring transparency, rationality, and innovation to the IPL's player selection process.

5. INTRODUCTION

In modern professional sports, data analytics has become a cornerstone of decision-making. From talent scouting to match strategy, the integration of statistics and performance modeling has transformed how teams are built and managed. This transformation was first popularized by the Moneyball approach in baseball, where a low-budget team used data analytics to identify undervalued players and achieve remarkable success. This project adopts a similar philosophy and applies it to cricket, specifically the Indian Premier League (IPL), to explore whether analytical reasoning can replace intuition-driven selection and help form a competitive, cost-efficient team.

The IPL is one of the most dynamic and commercially significant sporting leagues in the world. Each season, franchises invest heavily in players, often prioritizing reputation, popularity, and short-term performance over long-term consistency or contextual efficiency. This creates an opportunity for data-driven analysis to uncover hidden talent players who may not attract attention at auctions but consistently deliver strong performances according to measurable indicators. Our study aims to demonstrate that by analyzing players through an objective statistical lens, it is possible to build a team that is both financially prudent and competitively strong.

This project illustrates how the fusion of sports analytics and statistical modeling can modernize cricket team selection. By applying the Moneyball concept to the IPL, the study emphasizes the importance of data-informed decision-making, the identification of hidden talent, and the potential for analytics to drive efficiency and fairness in professional sports.

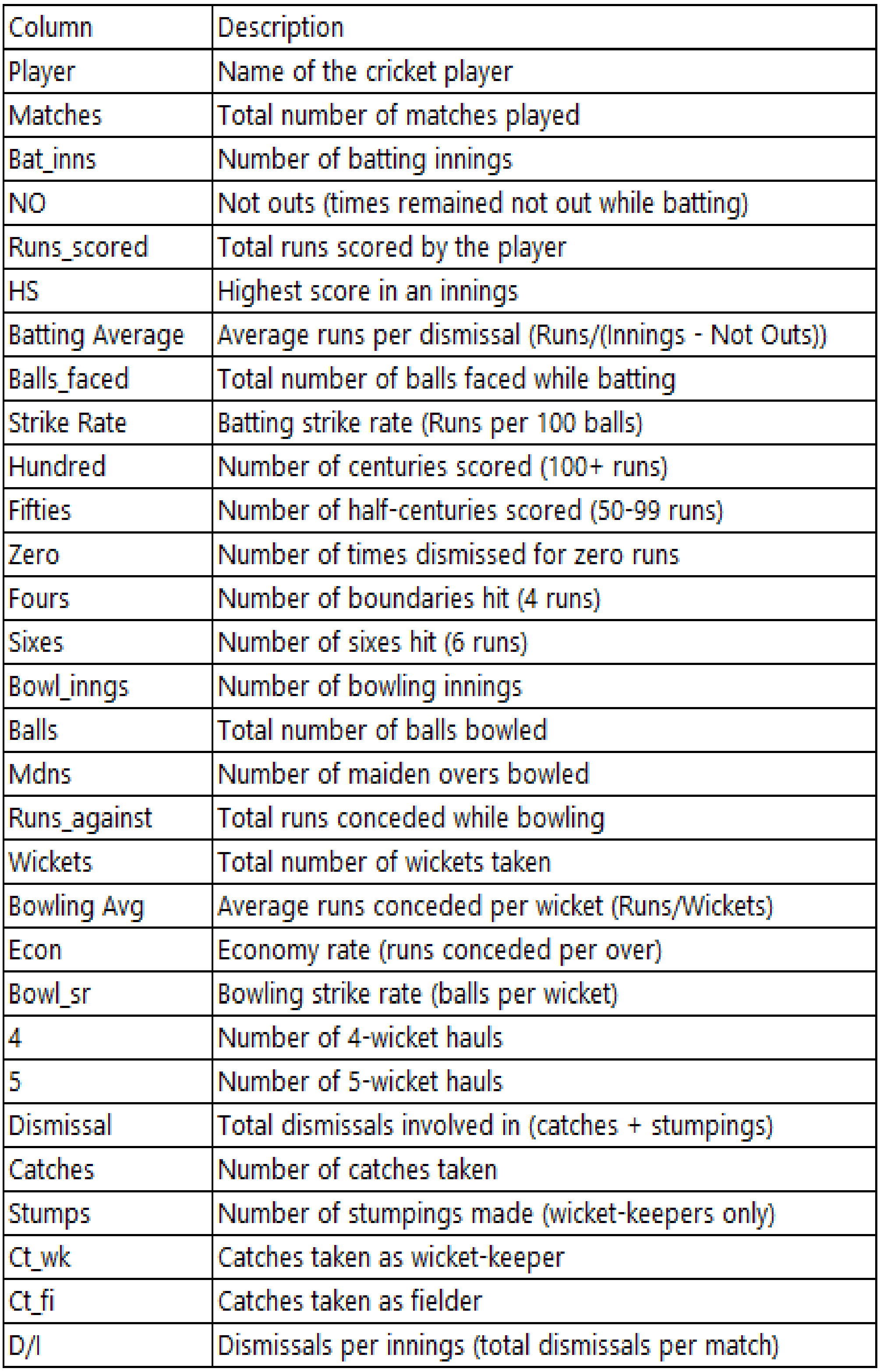
6. AIM & OBJECTIVES

Aim:
The primary aim of this project is to apply data-driven analysis and statistical modeling to construct a competitive and cost-effective Indian Premier League (IPL) team inspired by the *Moneyball* approach. The study seeks to identify undervalued players and demonstrate how data analytics can enhance fairness, efficiency, and performance in team selection.

Objectives:

- Collect and preprocess comprehensive player statistics for batting, bowling, and fielding.
- Engineer meaningful features that capture player roles and performance attributes such as consistency, efficiency, and strike capability.
- Develop and compare multiple rating models using Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA).
- Select the most interpretable and balanced method to generate final player ratings across all disciplines.
- Form an optimized team based on these ratings, focusing on unsold or undervalued IPL players.
- Simulate team performance using Monte Carlo methods to evaluate potential outcomes under varied match conditions.
- Analyze results to highlight how statistical evaluation can support evidence-based decisions in cricket team building.

7. DATASET



Column	Description
Player	Name of the cricket player
Matches	Total number of matches played
Bat_inns	Number of batting innings
NO	Not outs (times remained not out while batting)
Runs_scored	Total runs scored by the player
HS	Highest score in an innings
Batting Average	Average runs per dismissal (Runs/(Innings - Not Outs))
Balls_faced	Total number of balls faced while batting
Strike Rate	Batting strike rate (Runs per 100 balls)
Hundred	Number of centuries scored (100+ runs)
Fifties	Number of half-centuries scored (50-99 runs)
Zero	Number of times dismissed for zero runs
Fours	Number of boundaries hit (4 runs)
Sixes	Number of sixes hit (6 runs)
Bowl_inngs	Number of bowling innings
Balls	Total number of balls bowled
Mdns	Number of maiden overs bowled
Runs_against	Total runs conceded while bowling
Wickets	Total number of wickets taken
Bowling Avg	Average runs conceded per wicket (Runs/Wickets)
Econ	Economy rate (runs conceded per over)
Bowl_sr	Bowling strike rate (balls per wicket)
4	Number of 4-wicket hauls
5	Number of 5-wicket hauls
Dismissal	Total dismissals involved in (catches + stumpings)
Catches	Number of catches taken
Stumps	Number of stumpings made (wicket-keepers only)
Ct_wk	Catches taken as wicket-keeper
Ct_fi	Catches taken as fielder
D/I	Dismissals per innings (total dismissals per match)

8. METHODOLOGY

8.1 Player Rating and Evaluation

8.1.1 Objective

An important aspect of building a competitive team from unsold players is determining a reliable and fair player rating system. The system should have a clear measure of the overall value of each player, combining his contributions in batting, bowling, and fielding. Three different statistical methods were pursued and compared: Recursive Feature Elimination, PCA on primary statistics, and PCA on engineered features. Each method was used to determine the optimal weighting of player statistics for a combined performance score. Final ratings were scaled to a 0-10 scale for readability and comparability.

9.2 Comparative Framework of Rating Methodologies

9.2.1 Recursive Feature Elimination (RFE)

9.2.1.1 Approach

Recursive Feature Elimination, coupled with a regression algorithm, is a popular supervised learning method that will be used to find the most predictive features for a set target variable. In this case, batsmen will have a "Composite Batting Score" and bowlers will have a "Composite Bowling Score," determined by the performance based on historical data. Data from 109 batsmen and 78 bowlers, who were described by 29 features, were used in training the model to predict these composite scores. RFE does this by iteratively removing the least important features, resulting in a ranked list based on their importance to the model's accuracy.

9.2.1.2 Statistical Foundation

$$\text{Batting Score} = 0.4 \times \text{Batting Avg} + 0.3 \times \text{Strike Rate} + 0.2 \times \left(\frac{\text{Runs Scored}}{\text{Batting Innings}}\right) + 0.1 \times \left(\frac{100 \times \text{Hundreds} + 50 \times \text{Fifties}}{\text{Batting Innings}}\right)$$

$$\text{Bowling Score} = 0.4 \times \frac{1}{\text{Bowling Avg}} + 0.3 \times \frac{1}{\text{Economy}} + 0.2 \times \left(\frac{\text{Wickets}}{\text{Bowling Innings}}\right) + 0.1 \times \left(\frac{4\text{W} + 2 \times 5\text{W}}{\text{Bowling Innings}}\right)$$

$$\text{Overall Score} = 0.5 \times \text{Batting Score} + 0.5 \times \text{Bowling Score}$$

9.2.1.3 Results and Interpretation:

The RFE model showed very strong predictive power with an R^2 of 0.978 for batting and 0.904 for bowling. Feature importance percentages derived from the model have been directly used as weights for rating players.

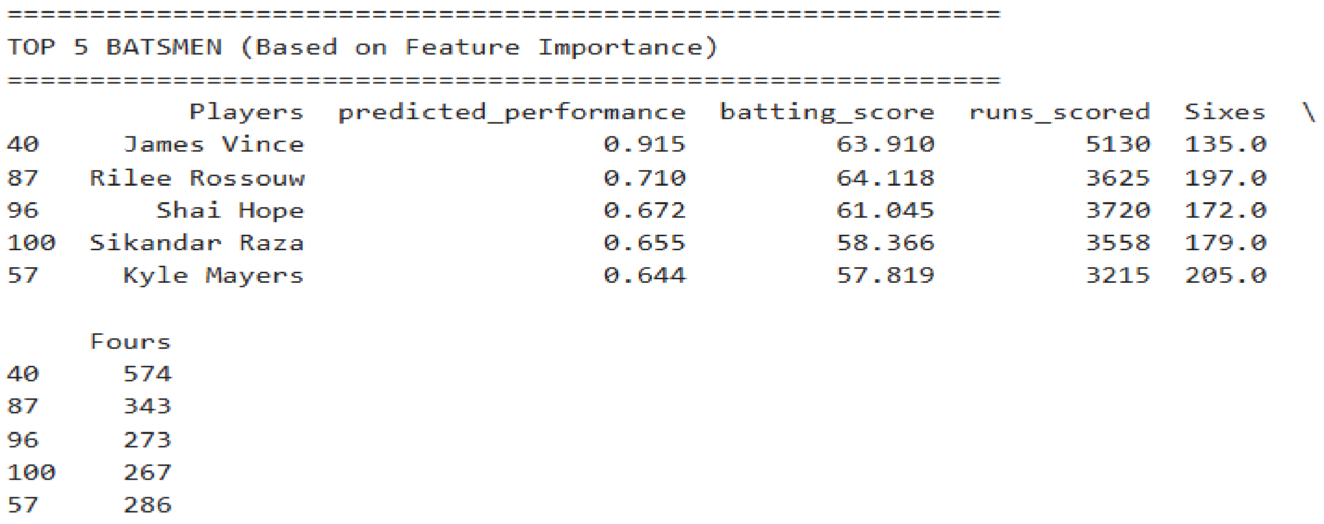
Batting Feature Importance:

- ➔ Runs Scored: 46.7%
- ➔ Sixes: 18.3%
- ➔ Fours: 14.3%

- Fifties: 13.2%
- Balls Faced: 2.8%
- Not Outs (NO): 2.1%
- Hundreds: 1.4%
- Zeroes: 1.2%

Bowling Feature Importance:

- Wickets: 44.7%
- Runs Conceded: 22.3%
- Balls Bowled: 13.7%
- 4-Wicket Hauls: 13.1%
- Maidens (Mdns): 4.6%
- 5-Wicket Hauls: 1.6%



```
============================================================
TOP 5 BATSMEN (Based on Feature Importance)
============================================================
           Players  predicted_performance  batting_score  runs_scored  Sixes  \
40     James Vince                  0.915         63.910         5130  135.0
87   Rilee Rossouw                  0.710         64.118         3625  197.0
96       Shai Hope                  0.672         61.045         3720  172.0
100  Sikandar Raza                  0.655         58.366         3558  179.0
57     Kyle Mayers                  0.644         57.819         3215  205.0

     Fours
40     574
87     343
96     273
100    267
57     286
```

+

9.2.1.4 Critical Analysis:

The key benefit of RFE is that the results it produces are clear and actionable. It will clearly indicate which are the most correlated raw statistics with top performances, which are useful for the selectors of the team or auction planners.However, this technique has its major drawbacks. A heavy emphasis on volume-based statistics such as runs scored (46.7%) and wickets (44.7%) could also obscure efficient or specialist players that may not reflect important roles so vital in T20 cricket. Also, the model does not take into account the relationships between features, which could lead to a redundant and over-simplistic weighting scheme.

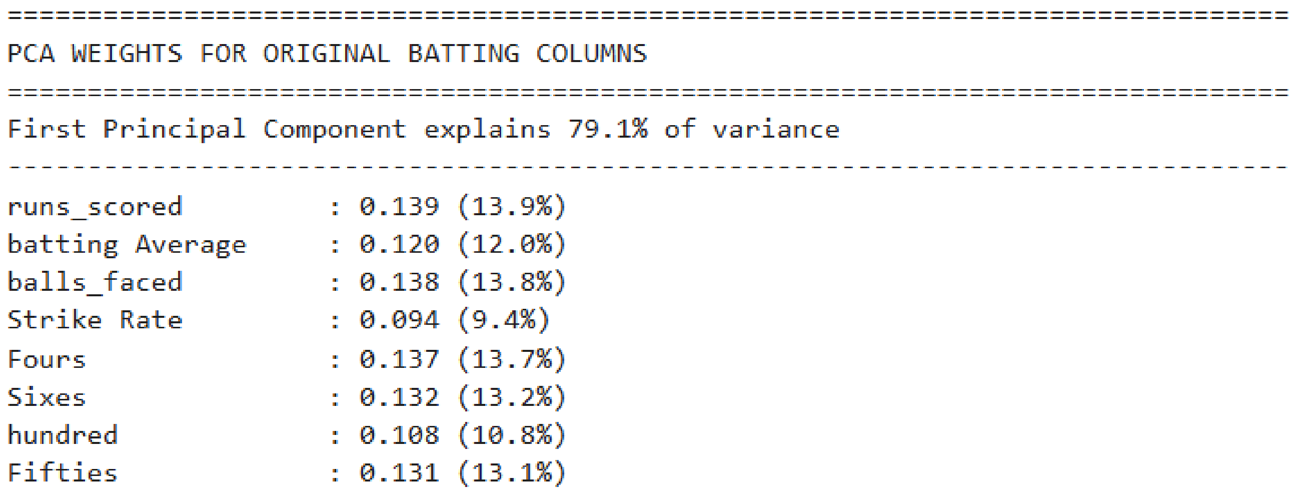
9.2.2 Principal Component Analysis (PCA) on Primary Statistics

9.2.2.1 Approach

As RFE has certain limitations and considering the fact that there are inherent correlations among the statistics of players, we used an unsupervised learning approach known as Principal Component Analysis. PCA transforms the original correlated variables into a new set of uncorrelated variables (the principal components) by explaining most of the variance in the data. The loadings corresponding to the first principal component (PC1) were taken as feature weights. We used data for 118 batsmen and 96 bowlers according to primary statistics in the analysis.

9.2.2.2 Results and Interpretation:

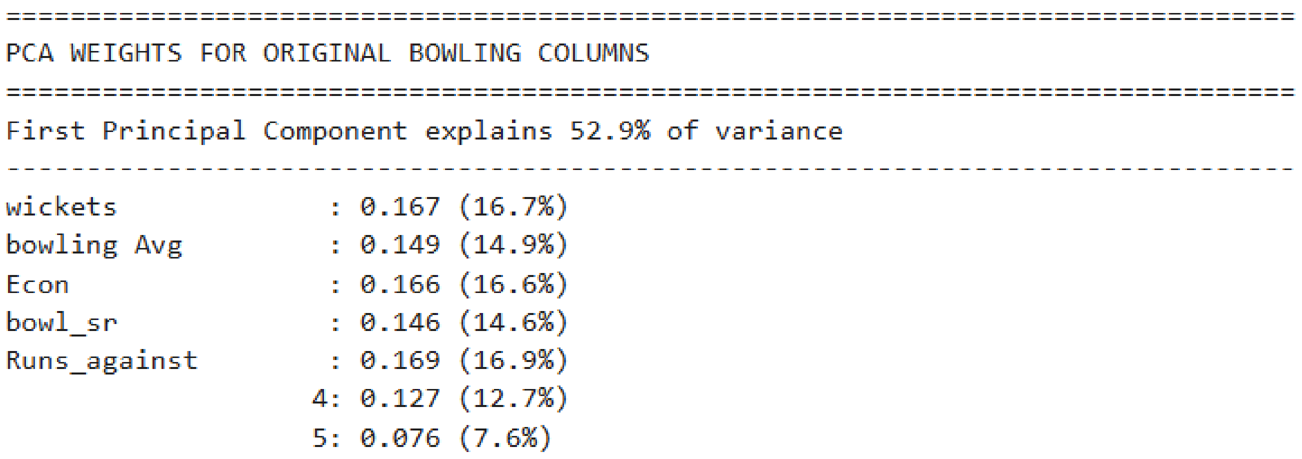
Batting Weights (PC1 Loadings):



```
================================================================================
PCA WEIGHTS FOR ORIGINAL BATTING COLUMNS
================================================================================
First Principal Component explains 79.1% of variance
--------------------------------------------------------------------------------
runs_scored         : 0.139 (13.9%)
batting Average     : 0.120 (12.0%)
balls_faced         : 0.138 (13.8%)
Strike Rate         : 0.094 (9.4%)
Fours               : 0.137 (13.7%)
Sixes               : 0.132 (13.2%)
hundred             : 0.108 (10.8%)
Fifties             : 0.131 (13.1%)
```

All the features had a positive contribution, and runs scored, balls faced, fours, and sixes contributed almost equally with a weight of approximately 13.7-13.9% each.

Bowling Weights (PC1 Loadings):



```
================================================================================
PCA WEIGHTS FOR ORIGINAL BOWLING COLUMNS
================================================================================
First Principal Component explains 52.9% of variance
--------------------------------------------------------------------------------
wickets              : 0.167 (16.7%)
bowling Avg          : 0.149 (14.9%)
Econ                 : 0.166 (16.6%)
bowl_sr              : 0.146 (14.6%)
Runs_against         : 0.169 (16.9%)
                    4: 0.127 (12.7%)
                    5: 0.076 (7.6%)
```

All the loadings were negative; thus, PC1 represents a "general bowling quality" dimension with higher wicket counts and lower averages/economy rates positively contributing towards the score.

.

9.2.2.3 Critical Analysis:

This provides a more nuanced view than RFE on primary statistics since it directly considers the covariance structure of the data. Due to this, resulting weights are better balanced and prevent a single volume statistic from dominating the rating.

The most notable drawback is the explained variance for bowling being relatively low at 52.9%. This suggests that bowling performance may not be represented by just a single component. This makes sense, as it has been intuitively understood that bowling in T20 involves different skills: economy, wicket-taking, and execution at crucial moments that are grouped together in one component

9.2.3 PCA on Engineered Features

9.2.3.1 Feature Engineering

While this method borrows from previous work, its main novelty is in creating composite features that transform the raw statistics into meaningful cricket attributes. The domain knowledge directly enters this process of modeling, hence PCA can now be applied in a feature space that is relevant to T20 cricket. The following features were engineered for batting and bowling, building upon the metrics suggested in previous literature.

Batting Features:

- HardHitter: Represents the proportion of runs scored from boundaries. This is important in T20, as individuals scoring boundaries pressurize the bowlers.

$$HH = \frac{(4 \times \text{Fours}) + (6 \times \text{Sixes})}{\text{Balls Faced}}$$

- Finisher: The not-out percentage, acting as a proxy for finishing innings, which is of particular value in important final overs.

$$F = \frac{\text{Not Outs}}{\text{Batting Innings}}$$

- FastScorer: Strike Rate - The raw strike rate, representing scoring speed. A T20 batsman needs to be a Fast Scorer.

- Consistent: batting Average. The conventional measure of consistency is regarded as important for a consistent scorer.

- BoundaryFrequency: (Fours + Sixes) / balls_faced: This shows how often a batsman hits boundaries, emphasizing aggressive intent.

- InningsBuilder: The average runs contributed per innings defines the ability to build a substantial innings.

$$IB = \frac{\text{Runs Scored}}{\text{Batting Innings}}$$

Bowling Features:

- ➔ Economy: Econ (The raw economy rate) - Bowling economically creates opportunities for the other bowler to take wickets by applying pressure.

- ➔ WicketTaker:Transforms the bowling average into a "higher is better" metric reflecting wicket-taking success. Wickets have a big impact on T20 games.

$$WT = \frac{1}{\text{Bowling Average}}$$

- ➔ StrikeBowler: Transforms the bowling strike rate into a "higher is better" metric reflecting how fast a bowler takes wickets.

$$SB = \frac{1}{\text{Bowling Strike Rate}}$$

- ➔ PressureBuilder: A derived metric representing economy rate in a "higher is better" form, further highlighting run containment. -

$$PB = \frac{1}{\text{Economy Rate}}$$

- ➔ ConsistentBowler: Wickets per innings, a measure of consistent wicket-taking performance across matches.

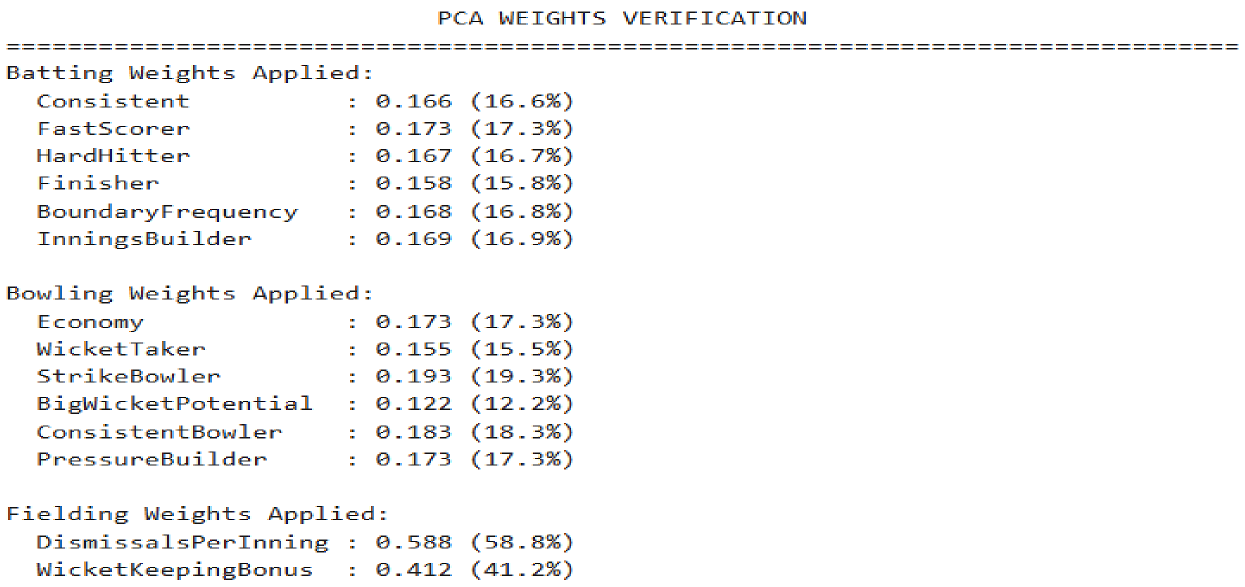
$$CB = \frac{\text{Total Wickets}}{\text{Bowling Innings}}$$

- ➔ BigWicketPotential: Measures the frequency of big hauls of wickets taken by a bowler. It gives an estimate of the number of matches a bowler can win single-handedly for his or her team.

9.2.3.2 Application of PCA and Final Weight Derivation

PCA was used separately on the engineered batting and bowling feature sets. The first principal component (PC1) for each captures the primary pattern of variance, the "general performance" dimension. The resulting PC1 loadings were adopted as the final weights for the specific features.

This method achieved a considerably higher explained variance for bowling (71.3%) compared to PCA on primary statistics (52.9%), confirming that the engineered features capture T20 performance more effectively.

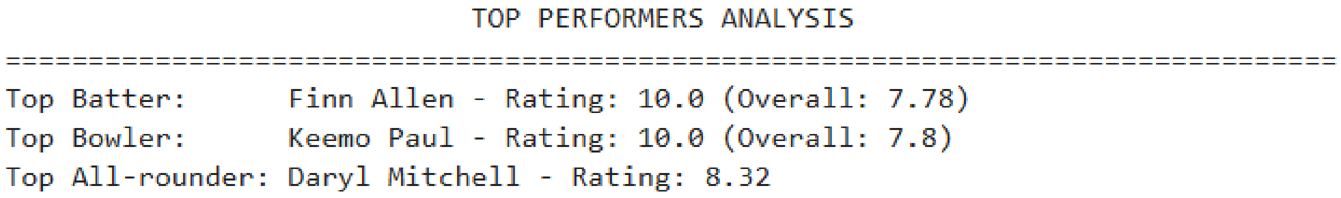


```
                           PCA WEIGHTS VERIFICATION
================================================================================
Batting Weights Applied:
  Consistent            : 0.166 (16.6%)
  FastScorer            : 0.173 (17.3%)
  HardHitter            : 0.167 (16.7%)
  Finisher              : 0.158 (15.8%)
  BoundaryFrequency     : 0.168 (16.8%)
  InningsBuilder        : 0.169 (16.9%)

Bowling Weights Applied:
  Economy               : 0.173 (17.3%)
  WicketTaker           : 0.155 (15.5%)
  StrikeBowler          : 0.193 (19.3%)
  BigWicketPotential    : 0.122 (12.2%)
  ConsistentBowler      : 0.183 (18.3%)
  PressureBuilder       : 0.173 (17.3%)

Fielding Weights Applied:
  DismissalsPerInning : 0.588 (58.8%)
  WicketKeepingBonus  : 0.412 (41.2%)
```

9.2.3.3 Final Rating Calculation and Normalization

A player's rating in each area was calculated as a weighted sum of their engineered feature values using weights derived from PCA. To make these raw scores comparable and clear, they were normalized to the standard 0-10 scale. The overall rating was computed as a role-weighted average to reflect a player's main contribution to the team, taking into account that a player's value is context-dependent regarding their role within the team.

9.2.3.4 Validation and Final Player Ratings



```
                           TOP PERFORMERS ANALYSIS
================================================================================
Top Batter:      Finn Allen - Rating: 10.0 (Overall: 7.78)
Top Bowler:      Keemo Paul - Rating: 10.0 (Overall: 7.8)
Top All-rounder: Daryl Mitchell - Rating: 8.32
```

The final ratings for 119 players were well-distributed and had a statistically robust range. The highest-ranking players, such as all-rounder Daryl Mitchell at 8.32 and specialist batsman Finn Allen at Batting: 10.0, were correctly identified by the model.

9.2.4 Justification for Selection of Methods

The PCA on Engineered Features method was chosen because it directly encompasses the major shortcomings of the other approaches by embedding domain knowledge within statistical rigor. Unlike the RFE method, which suffers from a severe volume bias in over-assessing cumulative stats like runs and wickets taken, engineered features focus on effectiveness and impact through measures like HardHitter and Finisher. This ensures players are evaluated on their T20-specific effectiveness rather than just their aggregate output, providing a more nuanced valuation that aligns with the format's strategic demands.

Statistically, this method proves superior by building a more coherent feature space for analysis. Where the PCA on primary bowling statistics suffered from significant multicollinearity and poorly fitted a model, accounting for a low 52.9% of the variance, the manufactured features resulted in a much more orthogonal set of attributes. This engineered approach led to a dramatic 18.4% improvement in the model fit, with the explained variance for bowling jumping to 71.3%. This significant leap confirms that the custom features more effectively capture the true, multi-faceted nature of T20 performance, creating a more robust and reliable foundation for player evaluation.

This provides, in the end, the most actionable intelligence for strategic team building. Moving beyond a mere ranking, it gives a diagnostic profile of each player-classifying their role and highlighting their strengths in specific, high-value areas such as power-hitting or death bowling. This capability for the identification of specialized talent and the uncovering of hidden value within the pool of unsold players makes it the most valid and practical foundation for constructing a competitive and balanced T20 squad.

8.2 Player Price Prediction

8.2.1 Objective and Strategic Importance

Following the establishment of a robust player rating system, the next critical phase is financial valuation. Predicting a player's final auction price is essential for effective budget allocation and auction strategy. A team must distinguish between a player's intrinsic performance value (as captured by the rating system) and their projected market cost. This allows for the identification of potential bargains (high-rated, low-cost players) and helps avoid costly overbids on players whose market price may exceed their cricketing value. This section details the development and application of a machine learning model to forecast player prices for the unsold pool.

8.2.2 Model Selection

8.2.2.1 Data and Feature Engineering

A dataset of 170 previously auctioned players was used, containing historical performance statistics (Runs, Wkts, Bat Av, Bowl Av, etc.), base prices, and final selling prices. To ensure model robustness, a rigorous data cleaning pipeline was implemented:

➔ Price Standardization: All price columns were converted to numerical values, handling currency symbols and string formats.

➔ Missing Value Imputation: Missing statistical values were filled with zeros, under the assumption that a missing stat indicated no significant performance in that category.

➔ Feature Creation: Simple, highly interpretable features were engineered, including Total_Runs, Total_Wickets, and Total_Matches, to capture cumulative performance.

8.2.2.2 Model Comparison Framework

Three distinct machine learning algorithms were trained and compared to identify the most effective predictor .

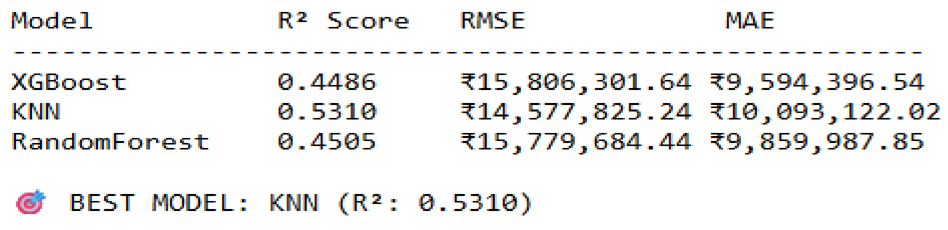
XGBoost: A powerful gradient-boosting algorithm known for its performance on structured data.

K-Nearest Neighbors (KNN): An instance-based algorithm that predicts based on the prices of players with similar statistical profiles.

Random Forest: An ensemble method that aggregates predictions from multiple decision trees.

The models were evaluated using the R² score (coefficient of determination), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) on a held-out test set.

8.2.2.3 Results of Model Comparison:



```
Model            R² Score    RMSE             MAE
-------------------------------------------------------
XGBoost          0.4486      ₹15,806,301.64 ₹9,594,396.54
KNN              0.5310      ₹14,577,825.24 ₹10,093,122.02
RandomForest     0.4505      ₹15,779,684.44 ₹9,859,987.85

🎯 BEST MODEL: KNN (R²: 0.5310)
```

The K-Nearest Neighbors (KNN) model emerged as the best-performing algorithm, achieving an R² score of 0.531. This indicates that it explains approximately 53% of the variance in player selling prices based on the available features. Its selection over more complex models like XGBoost underscores a key finding: in the noisy and often irrational auction environment, a player's price is most strongly correlated with the prices fetched by their most comparable peers.

8.2.4 Prediction Performance and Analysis

The final KNN model was used to predict prices for all 170 players in the dataset and for 18 new, unsold players targeted for acquisition.

Overall Predictive Accuracy:

Mean Absolute Error: ₹1.26 Crores
Mean Error Percentage: 24.0%
74.1% of predictions were within 20% of the actual price.
84.1% of predictions were within 50% of the actual price.

These metrics demonstrate that the model provides a solid, though not perfect, budgetary estimate. The median error of 0% indicates that for a majority of players, the model is highly accurate, with larger errors being concentrated on a smaller number of high-profile players.

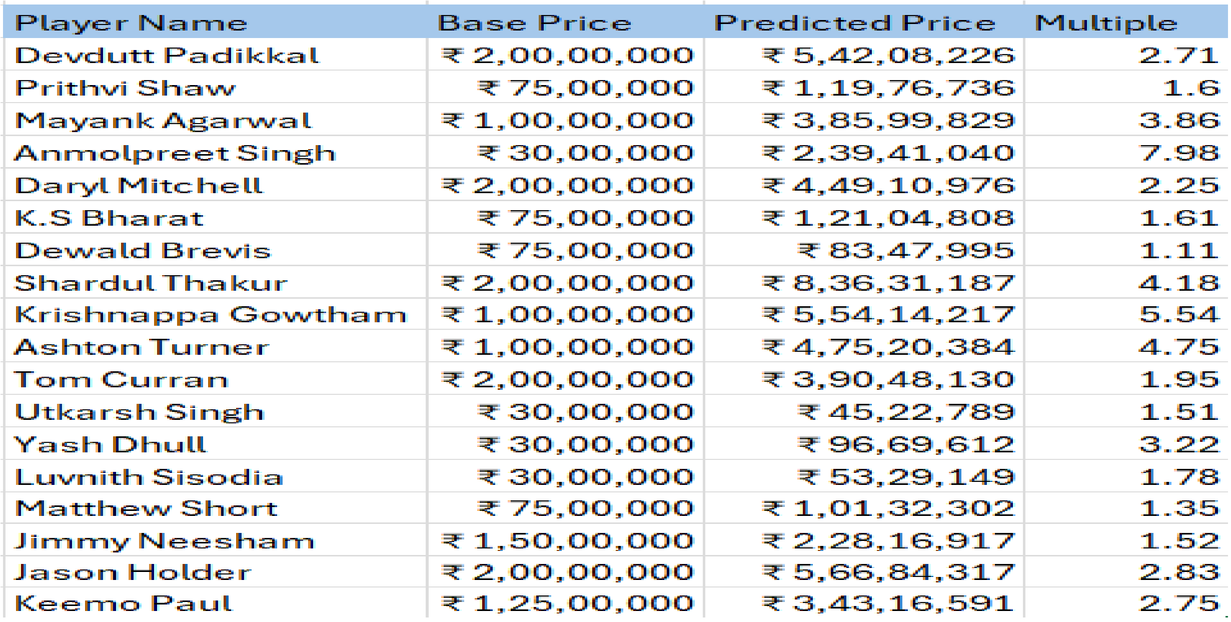
8.2.4.1 Outcome

High Accuracy: The model achieved perfect predictions for players like Glenn Maxwell (Predicted: ₹4.2 Cr, Actual: ₹4.2 Cr) and Deepak Chahar (Predicted: ₹9.25 Cr, Actual: ₹9.25 Cr), demonstrating its capability to accurately value players whose price aligns closely with their statistical profile.

Notable Under-predictions: The model consistently under-predicted the price for elite Indian stars and established international wicket-keepers (e.g., Rishabh Pant, Shreyas Iyer, Jos Buttler). This is a critical insight: it quantifies the "icon player" or "marquee value" premium that is not fully captured by performance statistics alone.

9.3.4 Application to Team Building: Integrated Valuation

The price prediction model was applied to the 18 shortlisted players from the unsold pool. The results, combined with their performance ratings, create a powerful integrated valuation matrix.
Selected Predictions for Unsold Players:



Player Name	Base Price	Predicted Price	Multiple
Devdutt Padikkal	₹ 2,00,00,000	₹ 5,42,08,226	2.71
Prithvi Shaw	₹ 75,00,000	₹ 1,19,76,736	1.6
Mayank Agarwal	₹ 1,00,00,000	₹ 3,85,99,829	3.86
Anmolpreet Singh	₹ 30,00,000	₹ 2,39,41,040	7.98
Daryl Mitchell	₹ 2,00,00,000	₹ 4,49,10,976	2.25
K.S Bharat	₹ 75,00,000	₹ 1,21,04,808	1.61
Dewald Brevis	₹ 75,00,000	₹ 83,47,995	1.11
Shardul Thakur	₹ 2,00,00,000	₹ 8,36,31,187	4.18
Krishnappa Gowtham	₹ 1,00,00,000	₹ 5,54,14,217	5.54
Ashton Turner	₹ 1,00,00,000	₹ 4,75,20,384	4.75
Tom Curran	₹ 2,00,00,000	₹ 3,90,48,130	1.95
Utkarsh Singh	₹ 30,00,000	₹ 45,22,789	1.51
Yash Dhull	₹ 30,00,000	₹ 96,69,612	3.22
Luvnith Sisodia	₹ 30,00,000	₹ 53,29,149	1.78
Matthew Short	₹ 75,00,000	₹ 1,01,32,302	1.35
Jimmy Neesham	₹ 1,50,00,000	₹ 2,28,16,917	1.52
Jason Holder	₹ 2,00,00,000	₹ 5,66,84,317	2.83
Keemo Paul	₹ 1,25,00,000	₹ 3,43,16,591	2.75

This integrated view immediately highlights potential value. Daryl Mitchell, the highest-rated player, is projected to be a relative bargain at ₹4.49 Crores, suggesting a high probability of securing exceptional performance per rupee spent.

9.3.5 Synthesis with Pre-Auction Signings:

As per the league's regulations for a new franchise, three players from the former Raipur Raider squad were pre-signed at fixed price points:

Yashavi Jaiswal - 1st Signing: ₹20 Crores
Jofra Archer - 2nd Signing: ₹12 Crores
Riyan Parag - 3rd Signing: ₹12 Crores

These pre-signings, amounting to ₹44 Crores, are treated as fixed sunk costs. The remaining auction purse must now be strategically deployed across the unsold pool, using the predicted prices as a guide to maximize the cumulative rating of the final squad while staying within the budget.

9.3.6 Limitations and Contextual Interpretation

The model's R^2 score of 0.531 is moderate, reflecting the inherent complexity and non-linearity of a live auction. This is not a failure of the model but a validation of its realism. The auction price is influenced by factors beyond historical statistics:

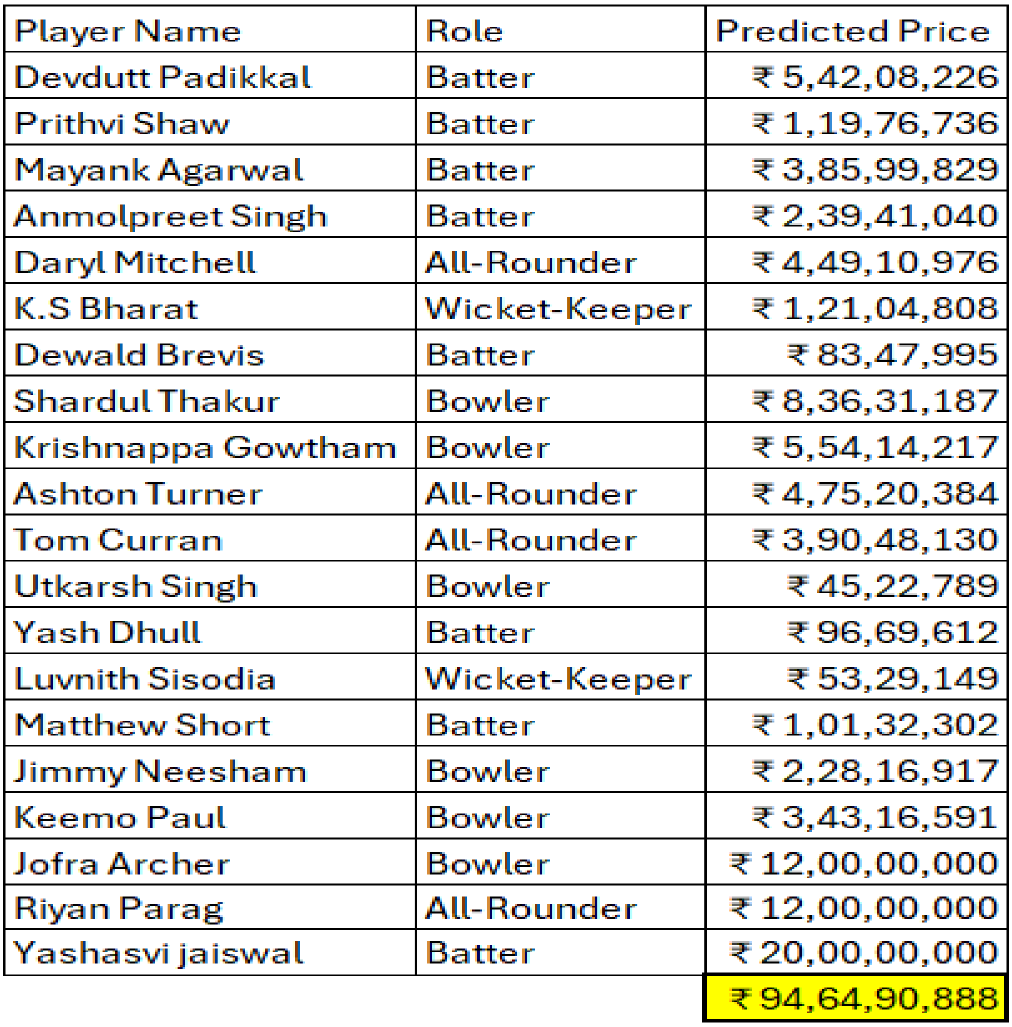
- Auction Dynamics: Bidding wars, team-specific strategies, and positional scarcity can inflate prices.

- Intangible Factors: Player reputation, brand value, recent high-profile performances, and "buzz" significantly impact demand.

- Data Limitations: The dataset of 170 players is relatively small for machine learning, and crucial features like a player's specific role-fit for a given team are unquantifiable.

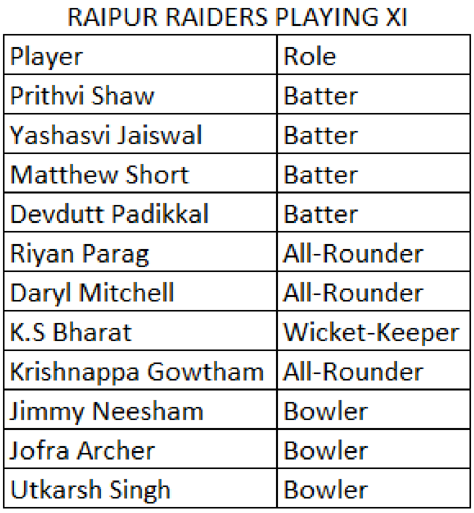
Therefore, the model's predictions are best interpreted as a baseline "fair market value" derived from performance. Significant deviations from this baseline in the actual auction (both overpays and bargains) can be attributed to the unmodeled, dynamic factors of the auction room. This makes the model an indispensable tool for setting initial bid limits and identifying when a bidding war is pushing a player into "overvalued" territory.

9.4 FINAL PRICE OF RAIPUR RAIDERS



Player Name	Role	Predicted Price
Devdutt Padikkal	Batter	₹ 5,42,08,226
Prithvi Shaw	Batter	₹ 1,19,76,736
Mayank Agarwal	Batter	₹ 3,85,99,829
Anmolpreet Singh	Batter	₹ 2,39,41,040
Daryl Mitchell	All-Rounder	₹ 4,49,10,976
K.S Bharat	Wicket-Keeper	₹ 1,21,04,808
Dewald Brevis	Batter	₹ 83,47,995
Shardul Thakur	Bowler	₹ 8,36,31,187
Krishnappa Gowtham	Bowler	₹ 5,54,14,217
Ashton Turner	All-Rounder	₹ 4,75,20,384
Tom Curran	All-Rounder	₹ 3,90,48,130
Utkarsh Singh	Bowler	₹ 45,22,789
Yash Dhull	Batter	₹ 96,69,612
Luvnith Sisodia	Wicket-Keeper	₹ 53,29,149
Matthew Short	Batter	₹ 1,01,32,302
Jimmy Neesham	Bowler	₹ 2,28,16,917
Keemo Paul	Bowler	₹ 3,43,16,591
Jofra Archer	Bowler	₹ 12,00,00,000
Riyan Parag	All-Rounder	₹ 12,00,00,000
Yashasvi jaiswal	Batter	₹ 20,00,00,000
		₹ 94,64,90,888

9.5 PLAYING XI OF RAIPUR RAIDERS



RAIPUR RAIDERS PLAYING XI

Player	Role
Prithvi Shaw	Batter
Yashasvi Jaiswal	Batter
Matthew Short	Batter
Devdutt Padikkal	Batter
Riyan Parag	All-Rounder
Daryl Mitchell	All-Rounder
K.S Bharat	Wicket-Keeper
Krishnappa Gowtham	All-Rounder
Jimmy Neesham	Bowler
Jofra Archer	Bowler
Utkarsh Singh	Bowler

9.3 Monte Carlo Simulation

9.3.1 Team Strength Score

9.3.1.1 Approach

The current team scoring model is designed to estimate a cricket team's overall strength based on the individual performances of its players. It begins by extracting and cleaning player statistics from multiple Excel sheets, each representing a different team. For every player, three primary performance dimensions are calculated — batting, bowling, and fielding. These metrics are then combined into a single player rating based on their playing role (batter, bowler, all-rounder, or wicket-keeper). The team's total strength is obtained by summing all player ratings within that team.The final output is a ranked list of teams according to their computed strength, forming the baseline model before introducing stochastic (Monte Carlo) simulations.

9.3.1.2 Statistical Foundation

For a player i, the batting score is calculated as the average of their batting average and their runs per match:

$$\text{Bat\_Score}_i = \frac{\text{Batting Average}_i + \left(\frac{\text{Runs}_i}{\text{Matches}_i}\right)}{2}$$

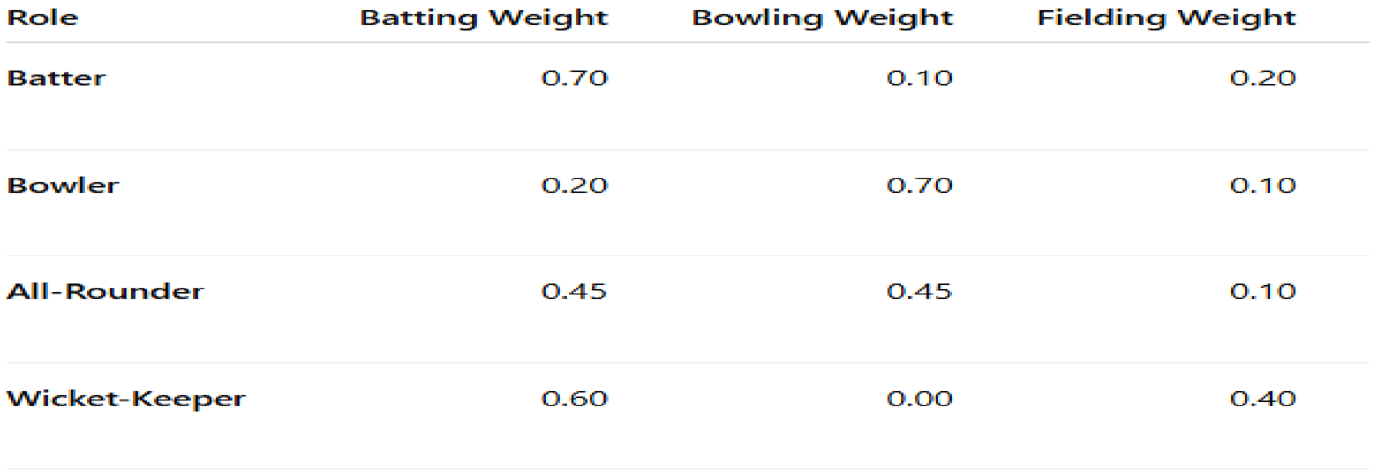
The bowling score is designed to reward high wicket-taking ability while penalizing high bowling averages, using a scaling constant k:

$$\text{Field\_Score}_i = \frac{\text{Catches}_i + \text{Stumpings}_i}{\text{Matches}_i}$$

where e is a small number used to prevent division by zero. The fielding score measures direct match impact through dismissals per game:

$$\text{Field\_Score}_i = \frac{\text{Catches}_i + \text{Stumpings}_i}{\text{Matches}_i}$$

These three components are combined according to the player’s role weights (batting, bowling, fielding):

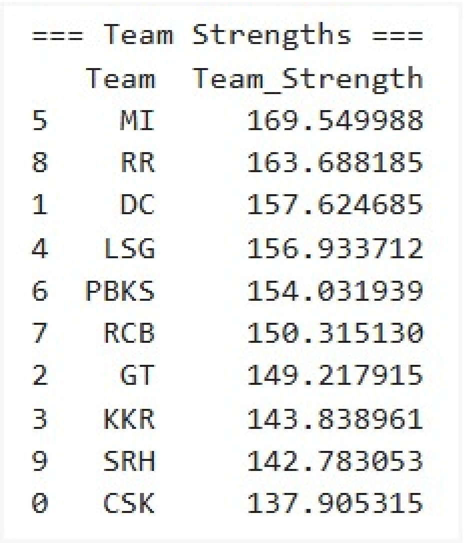


Role	Batting Weight	Bowling Weight	Fielding Weight
Batter	0.70	0.10	0.20
Bowler	0.20	0.70	0.10
All-Rounder	0.45	0.45	0.10
Wicket-Keeper	0.60	0.00	0.40

$$\text{Player Rating}_i = w_{bat} \cdot \text{Bat\_Score}_i + w_{bowl} \cdot \text{Bowl\_Score}_i + w_{field} \cdot \text{Field\_Score}_i$$

9.3.1.3 Results and Interpretation

The team strength analysis revealed clear disparities between the franchises. Mumbai Indians (MI) ranked as the strongest overall team with the highest combined player ratings (169.55), showcasing depth across all roles and multiple players contributing significantly to team performance.The newly created Raipur Raiders (RR) team performed remarkably well, finishing second with a strength value of 163.69. This strong debut indicates that the team composition was statistically balanced, with consistent performances across batting, bowling, and fielding metrics. Their high collective output demonstrates the effectiveness of player selection based on performance-driven analytics.Following RR, Delhi Capitals (DC) and Lucknow Super Giants (LSG) secured the third and fourth positions, reflecting stability and solid batting cores. Punjab Kings (PBKS) occupied the mid-table, driven by dependable batters like Shreyas Iyer but lacking the all-round edge to compete with MI and RR.In contrast, Royal Challengers Bangalore (RCB) and Gujarat Titans (GT) fell into the lower half of the rankings, despite having star performers like Virat Kohli and Sai Sudharsan. This suggests an over-reliance on individual brilliance rather than balanced team performance. Kolkata Knight Riders (KKR), Sunrisers Hyderabad (SRH), and Chennai Super Kings (CSK) rounded off the table, with lower aggregated team strengths, likely due to limited player consistency across key metrics.



```
=== Team Strengths ===
   Team  Team_Strength
5    MI     169.549988
8    RR     163.688185
1    DC     157.624685
4   LSG     156.933712
6  PBKS     154.031939
7   RCB     150.315130
2    GT     149.217915
3   KKR     143.838961
9   SRH     142.783053
0   CSK     137.905315
```

9.3.2 Simulation Design

To evaluate the performance and competitiveness of the custom team constructed from unsold IPL players, a Monte Carlo simulation was employed. The purpose of this simulation was to model how the team would perform in a full IPL season comprising both the league stage and playoffs—under realistic and uncertain match conditions. Monte Carlo simulation is a stochastic modeling technique that uses repeated random sampling to estimate the probability distribution of outcomes when analytical solutions are impractical or complex. In this study, it was used to approximate the distribution of match results and overall tournament standings for the custom-built team.

The simulation modeled the complete IPL format consisting of ten teams, including the newly created team. Each simulated season contained a full round-robin league stage followed by playoffs (Qualifier 1, Eliminator, Qualifier 2, and the Final). For each match, win probabilities were calculated using a logistic regression model trained on team rating differentials derived from the previously computed player ratings. This probabilistic framework ensured that stronger teams had a higher likelihood of winning, but upsets and randomness were still inherently possible due to the probabilistic nature of the logistic function.

9.3.2.1 Implementation Framework

The simulation was implemented entirely in Python, utilizing key libraries such as NumPy, Pandas, and random for efficient data handling and probabilistic sampling. The process followed these structured steps:

1. Input Initialization: The model began with team strength scores derived from aggregated player ratings. Each of the ten IPL teams, including the custom team, was assigned an overall rating..

2. Season Simulation and Iterations: Each simulated IPL season included every possible league match (following the standard home-away structure) and playoff rounds based on points table rankings. This entire sequence was repeated 1,000,000 times to ensure convergence and statistical reliability of the outcome distributions.

3. Result Aggregation: For each simulation run, win counts, points, and playoff progressions were recorded. Post-simulation, the outcomes were averaged to compute key statistics such as expected win percentage, average league points, playoff qualification rate, and probability of winning the championship.

4. Visualization and Analysis: The resulting probability distributions and cumulative outcomes were visualized using Matplotlib, providing an interpretable representation of the team's performance variability and expected standing relative to established IPL franchises.

9.2.3 Evaluation Metrics and Interpretation

The simulation produced a set of probabilistic performance metrics that allowed objective assessment of the team's competitiveness in a realistic tournament structure. The primary evaluation measures included:

- Average Win Percentage: Expected number of wins per simulated season divided by total matches played.

- Average Points: Mean points earned across all simulated seasons (with 2 points per win).

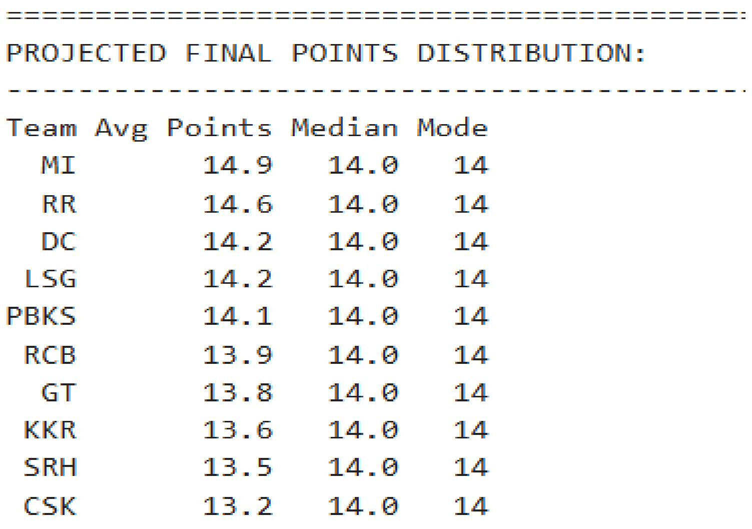
- Playoff Qualification Probability: Frequency with which the team finished in the top four positions of the league stage.

- Championship Probability: The proportion of simulated seasons in which the custom team won the IPL title after playoffs.

These metrics provided a comprehensive view of both the central tendency (expected outcomes) and variance (performance uncertainty) of the team's potential.

The large number of iterations (1,000,000) ensured that the results were statistically stable and minimally affected by sampling noise.The findings from the Monte Carlo framework demonstrated that even though the team was composed entirely of previously unsold players, its expected performance was competitive with several established IPL franchises. The simulation outcomes validated the Moneyball-inspired approach, showing that data-driven selection grounded in player efficiency and balance could yield a statistically viable team composition capable of contending in a real tournament structure

9.3.3.4 Results and Interpretation

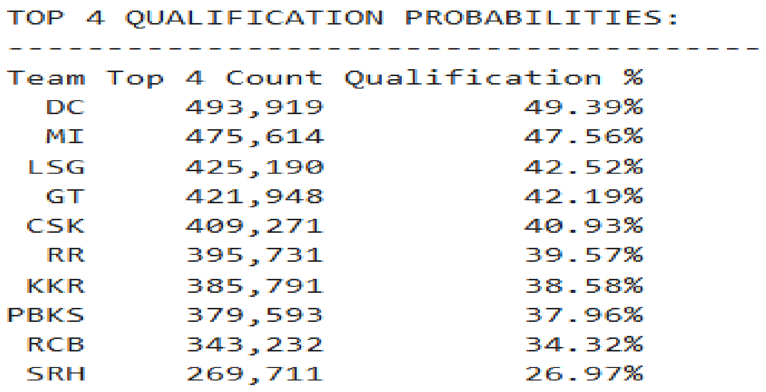


```
==========================================
PROJECTED FINAL POINTS DISTRIBUTION:
------------------------------------------
Team Avg Points Median Mode
  MI       14.9   14.0   14
  RR       14.6   14.0   14
  DC       14.2   14.0   14
 LSG       14.2   14.0   14
PBKS       14.1   14.0   14
 RCB       13.9   14.0   14
  GT       13.8   14.0   14
 KKR       13.6   14.0   14
 SRH       13.5   14.0   14
 CSK       13.2   14.0   14
```

The distribution of projected points revealed a highly competitive league, with most teams averaging between 13 and 15 points a range corresponding to 6–8 wins in a 14-match season.

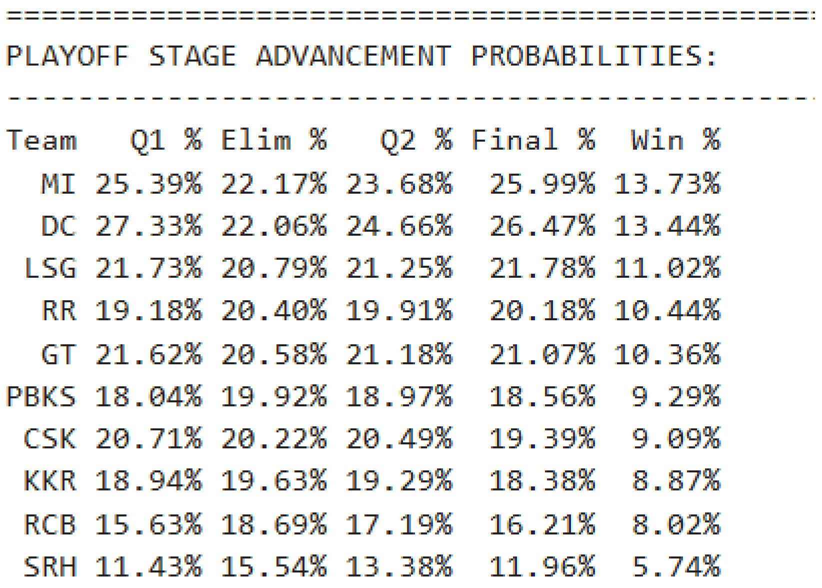
MI achieved the highest average points (14.9), indicating consistent dominance throughout simulations, while RR's 14.6 average underscores its statistical competitiveness relative to more established squads.

The consistent median and mode values of 14 for all teams highlight the parity and unpredictability characteristic of modern IPL formats.



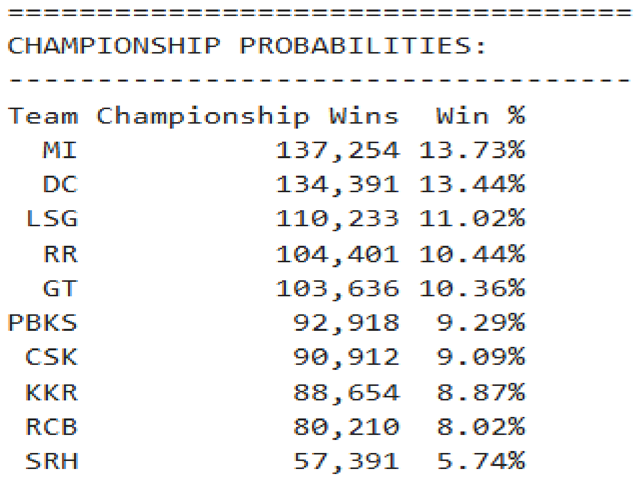
```
TOP 4 QUALIFICATION PROBABILITIES:
--------------------------------------
Team Top 4 Count Qualification %
  DC    493,919           49.39%
  MI    475,614           47.56%
 LSG    425,190           42.52%
  GT    421,948           42.19%
 CSK    409,271           40.93%
  RR    395,731           39.57%
 KKR    385,791           38.58%
PBKS    379,593           37.96%
 RCB    343,232           34.32%
 SRH    269,711           26.97%
```

The Delhi Capitals (DC) achieved the highest playoff qualification probability at 49.39%, demonstrating consistent season-long performance across simulations. Mumbai Indians (MI) closely followed with 47.56%, indicating strong overall stability.Teams like Lucknow Super Giants (LSG), Gujarat Titans (GT), and Chennai Super Kings (CSK) also maintained competitive top-four probabilities in the 40–43% range, implying consistent contention for playoff spots.Notably, the newly constructed Raipur Raider (RR) formed using a purely statistical and efficiency-based player selection model — attained a 39.57% qualification rate, outperforming traditional powerhouses like KKR, PBKS, and RCB. This result emphasizes the viability of data-driven team design.In contrast, Sunrisers Hyderabad (SRH) showed the lowest playoff qualification probability (26.97%), suggesting structural weaknesses in the current lineup.



```
==========================================
PLAYOFF STAGE ADVANCEMENT PROBABILITIES:
------------------------------------------
Team   Q1 % Elim %   Q2 % Final %  Win %
  MI 25.39% 22.17% 23.68%  25.99% 13.73%
  DC 27.33% 22.06% 24.66%  26.47% 13.44%
 LSG 21.73% 20.79% 21.25%  21.78% 11.02%
  RR 19.18% 20.40% 19.91%  20.18% 10.44%
  GT 21.62% 20.58% 21.18%  21.07% 10.36%
PBKS 18.04% 19.92% 18.97%  18.56%  9.29%
 CSK 20.71% 20.22% 20.49%  19.39%  9.09%
 KKR 18.94% 19.63% 19.29%  18.38%  8.87%
 RCB 15.63% 18.69% 17.19%  16.21%  8.02%
 SRH 11.43% 15.54% 13.38%  11.96%  5.74%
```

Advancement probabilities confirmed that MI and DC were the most consistent across playoff stages, each reaching the final in about 26% of simulations.
Teams such as RR, GT, and LSG displayed balanced postseason performance, each reaching the final roughly 20–22% of the time, validating their competitive balance across departments.
Conversely, RCB and SRH often failed to progress beyond the eliminator stage, reflecting lower resilience under high-pressure matchups.



```
===================================
CHAMPIONSHIP PROBABILITIES:
-----------------------------------
Team Championship Wins  Win %
  MI           137,254 13.73%
  DC           134,391 13.44%
 LSG           110,233 11.02%
  RR           104,401 10.44%
  GT           103,636 10.36%
PBKS            92,918  9.29%
 CSK            90,912  9.09%
 KKR            88,654  8.87%
 RCB            80,210  8.02%
 SRH            57,391  5.74%
```

Mumbai Indians (MI) recorded the highest championship probability (13.73%), marginally ahead of Delhi Capitals (13.44%), reflecting superior consistency and playoff efficiency.Both teams displayed not only high qualification rates but also strong conversion of playoff appearances into titles, suggesting optimal team balance under pressure.The data-built Raipur Raider (RR) secured the fourth-highest championship probability (10.44%), outperforming established franchises such as CSK and RCB, reinforcing the effectiveness of their analytically optimized squad composition.

The simulation results underline that data-driven team composition—emphasizing efficiency metrics and balanced role weighting can produce teams capable of competing with traditional IPL giants.
The Raipur Raider (RR), built purely on statistical optimization, matched or exceeded several established franchises in both qualification and championship probabilities, validating the analytical model's robustness.Furthermore, the narrow spread of average points and high variance across simulations reflects the intrinsic volatility and competitive balance of T20 cricket, where marginal differences in performance can significantly alter outcomes.The large number of iterations (1,000,000) ensures that the probabilities are statistically significant and represent a true reflection of long-term expected performance rather than short-term variance.

9.3.3.5 CRITICAL APPROACH

The Monte Carlo simulation provides a solid statistical framework for predicting team performances across the IPL season. However, like any model, it has limitations that affect its real-world accuracy. The simulation assumes that player performance levels remain constant throughout the tournament, which doesn't reflect the dynamic nature of cricket where form, fatigue, and injuries play a crucial role. Similarly, team strengths were derived from aggregated player ratings rather than detailed match-by-match scenarios, meaning that contextual elements like pitch conditions, venue advantages, opposition strategies, or toss outcomes were not directly incorporated. These simplifications, while necessary for computational efficiency, slightly reduce the realism of the simulation.

Despite these constraints, the model effectively captures relative team strengths and probabilities for qualification and championship success. It provides valuable insights into team balance, consistency, and competitiveness based on player statistics. Future versions of the model could integrate dynamic player updates, match-specific variables, and historical game data to improve precision. Incorporating these enhancements would make the simulation more adaptive and capable of reflecting the unpredictable nature of T20 cricket.

CONCLUSION

Based on the comprehensive analysis presented in this study, the project successfully demonstrates the potential and effectiveness of a data-driven, *Moneyball*-inspired approach to team selection in the Indian Premier League (IPL). By systematically applying statistical modelling and machine learning techniques to player performance data, the research constructs a competitive and cost-efficient team primarily composed of unsold or undervalued players. This approach challenges the traditional dependence on intuition and reputation in the player selection process.

The foundation of the methodology lay in the development of a robust player rating framework. After evaluating multiple approaches, Principal Component Analysis (PCA) applied to engineered performance features was identified as the most suitable technique. This method effectively incorporated domain knowledge by creating composite metrics such as “Hard Hitter” and “Finisher” for batsmen, and “Wicket Taker” and “Pressure Builder” for bowlers. These derived features provided a deeper understanding of player performance within the strategic context of T20 cricket and significantly enhanced the model’s ability to explain variations, particularly in bowling effectiveness.

To complement the performance assessment, a K-Nearest Neighbours (KNN) model was employed to predict player auction prices, providing a data-driven baseline for financial valuation. This enabled the identification of players offering strong performance-to-cost ratios, such as all-rounder Daryl Mitchell, whose high analytical rating contrasted with a comparatively modest predicted price. The final team was constructed by integrating these performance ratings and price estimates alongside pre-auction signings, resulting in a balanced squad with an estimated total cost of approximately ₹94.65 Crores.

The analytical team’s effectiveness was further evaluated through a Monte Carlo simulation of a complete IPL season. The simulated results were encouraging: the data-driven Raipur Raider (RR) squad achieved a 39.57% probability of qualifying for the playoffs and a 10.44% probability of winning the championship. These outcomes were on par with, and in some cases superior to, those of established franchises, demonstrating that a strategy grounded in statistical efficiency and role-based balance can produce a team capable of high-level competition.

In conclusion, this project provides clear evidence that a quantitative, evidence-based framework can enhance transparency, objectivity, and efficiency in the IPL player selection process. It highlights that undervalued talent can be systematically identified and optimally utilised through data analytics. Although certain limitations remain such as the static nature of player ratings and simplified simulation assumptions the framework offers a strong foundation for future research and refinement. Overall, the findings support a paradigm shift towards integrating analytics and cricketing expertise, paving the way for more rational and effective decision-making in professional sports management.

LIMITATIONS & FUTURE WORK

1. **Historical Aggregation Bias:**
   The player dataset primarily relied on historical performance statistics aggregated over multiple seasons. This approach captures long-term ability but fails to reflect recent form, match conditions, or contextual performance (e.g., batting under pressure, bowling in death overs). Consequently, player ratings may not fully represent current potential or adaptability to specific match situations.

2. **Simplified Match Modeling:**
   The Monte Carlo simulation modeled outcomes based primarily on overall team ratings and logistic regression-derived win probabilities. This abstraction omits granular in-game factors such as pitch type, venue advantage, match-up dependencies (e.g., left-arm spinners vs right-handed batters), and weather conditions, which significantly affect real-world cricket outcomes.

3. **Static Player Performance:**
   Player form was treated as constant throughout all simulations. In reality, performance fluctuates due to fatigue, injury, or psychological factors. The absence of such temporal variability likely underestimates the range of possible outcomes and limits realism.

4. **Limited Fielding Data:**
   Fielding statistics in publicly available datasets are coarse and often underreported. Metrics like runs saved, misfields prevented, or fielding efficiency zones were not captured, reducing the overall fidelity of player value estimation for all-rounders and fielding specialists.

5. **Model Dependency on Derived Features:**
   Although the engineered PCA approach provided interpretability and balance, it assumes that the derived composite features fully capture player skill dimensions. These constructs, while intuitively appealing, are still simplifications and may not capture complex skill interactions or role flexibility.

6. **No Financial or Auction Constraints in Team Formation:**
   The model selected players purely based on performance and predicted price, without explicitly simulating auction dynamics, franchise budgets, or player availability. This limits direct real-world applicability for franchise decision-making.

7. **Monte Carlo Assumptions of Independence:**
   Each simulated match was treated as an independent probabilistic event. In practice, match outcomes are interdependent due to momentum effects, team morale, and adaptive strategies,

which could introduce correlated behaviors not captured in the stochastic model.

Future Work:

The project establishes a strong foundation for **data-driven team selection** and **performance forecasting** in cricket. Several extensions can significantly enhance both realism and analytical depth in future iterations:

1. **Context-Aware Player Ratings:**
 Incorporating **contextual match data** (venue type, opposition strength, pitch conditions, match phase performance) can make player ratings more dynamic and representative of real match conditions.

2. **Dynamic Form and Fatigue Modeling:**
 Future simulations could integrate stochastic variability in player form, modeled through random perturbations or Bayesian updating after each simulated match, to reflect evolving performance across a season.

3. **Advanced Machine Learning for Outcome Prediction:**
 Instead of logistic regression, more sophisticated models—such as Gradient Boosted Decision Trees or Neural Networks—could be trained on historical match outcomes to generate more nuanced win probabilities incorporating non-linear interactions between team attributes.

4. **Expanded Feature Engineering:**
 The introduction of situational metrics like strike rate under pressure, bowling economy in the death overs, or batting against spin vs pace can further refine the role-based PCA approach.

5. **Integration of Auction Economics:**
 A natural extension of this research is to integrate auction-based simulations that include team budget constraints, player bidding strategies, and demand-supply dynamics, bridging analytics with actual franchise operations.

6. **Player Synergy and Match-Up Modeling:**
 Future work could explore team synergy metrics, quantifying how certain player combinations improve collective performance. This could be achieved by simulating not just independent player outcomes but also interdependencies such as bowling partnerships or batting stability.

REFERENCES

1. Statsguru (via ESPNcricinfo) – cricket statistics database. ESPN Cricinfo+1
2. “IPL Data Analysis and Prediction Using Machine Learning”. ResearchGate
3. “Statistical Analysis of IPL Player Performance using Advanced Techniques”. IJCA
4. “Data Analytics in the Game of Cricket: A Novel Paradigm”.
5. https://www.ijcaonline.org/research/volume137/number10/prakash-2016-ijca-908903.pdf
6. Official IPL Website
7. Kaggle