2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS) | 979-8-3315-3925-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/WorldSUAS66815.2025.11199299

Sports Analytics for IPL Auctions Machine Learning-Based Player Valuation Using Performance Data

Mr. Lokesh Sahu[1]
School of CSIT
Symbiosis University of Applied Sciences
Indore, India
lokesh.sahu@suas.ac.in

Dr. Durgesh Mishra[2]
School of CSIT
Symbiosis University of Applied Sciences
Indore, India
durgeshmishra@suas.ac.in

Rohit Manna[3]
School of CSIT
Symbiosis University of Applied Sciences
Indore, India
rohitmanna55@gmail.com

Abstract –This research presents a predictive model developed in R to estimate the auction prices of Indian Premier League (IPL) players, based on their performance data and roles. Key performance indicators, such as runs, strike rate, wickets, and economy rate, were analyzed to uncover trends linking player statistics to their market value. The model aims to reduce biases in auction decisions by providing a data-driven approach, offering franchises a more strategic tool for decision-making. This study highlights the integration of machine learning techniques in sports analytics to improve the accuracy and fairness of player valuations during IPL auctions.

Keywords- Predictive model, IPL auction, Performance data, Machine learning, Player valuation

INTRODUCTION

The Indian Premier League (IPL) is one of the most commercially successful and widely followed cricket tournaments globally. Player auction prices are influenced by both measurable performance data and subjective preferences of team management, often resulting in inconsistencies in valuation. [1] [2] Despite the availability of rich historical performance data, there remains a significant gap in applying machine learning techniques to accurately predict player auction prices. Leveraging data science in this context can provide teams with a more objective, transparent, and strategic framework for evaluating player worth, ultimately enhancing decision-making during auctions. [3]

This research aims to address the challenge of accurately estimating IPL player auction prices using a data-driven ensemble learning techniques like Random Forest and Gradient Boosting Machines (GBM) have demonstrated effectiveness in numerical prediction tasks, their application in auction price estimation remains underexplored. [10] [11] This project addresses that gap by developing a predictive model using measurable player approach. The project involves cleaning and preprocessing IPL player data, performing exploratory data analysis (EDA) to identify key trends, and training various regression models in R to predict auction values. [4] By evaluating the performance of these models, the most reliable one will be selected to support consistent player valuation. This approach minimizes subjective bias and enhances the strategic planning capabilities of franchises, offering a practical application of machine learning in sports analytics.

LITERATURE REVIEW

Sports analytics has seen significant advancements in recent years, with research primarily focused on predicting player performance, match outcomes, and optimizing team strategies. Statistical and machine learning techniques have been widely adopted to assess batting and bowling averages, forecast win probabilities, and evaluate player efficiency. The adoption of sabermetrics from baseball into cricket—often referred to as Criclytics—has provided a foundation for performance-based evaluation in the T20 format. [5] [6] Prior studies have explored player ranking systems, form analysis, and fantasy point predictions using regression and classification algorithms such as Support Vector Machines (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN). Emerging trends include the integration of real-time data streams, sentiment analysis from social media, and video-based performance tracking. [7]

Despite these advancements, limited research has been conducted on predicting IPL player auction prices, which involves not only performance metrics but also the dynamics of market demand, player roles, and franchise financial strategies. [8] [9] While regression models and performance indicators to estimate auction prices in a more systematic and objective manner. [12]

METHODLOGY

The methodology follows a structured pipeline covering data collection, preprocessing, feature engineering, and model evaluation to ensure prediction accuracy. Visualizations such as scatter plots, box plots, and a correlation heatmap aided in identifying key patterns and relationships. These graphical tools supported both analysis and model refinement. The complete workflow is summarized in Figure 1.

- **Data Collection**
  The dataset used for this project was created manually by collecting player statistics from the official IPL website and ESPNcricinfo. It contains 199 records and 14 attributes, including both categorical (e.g., playing role, batting style) and numerical variables (e.g., runs scored, wickets taken, batting average).
- **Data Cleaning & Preprocessing**
  The dataset was first cleaned by removing duplicate entries. Missing values were then imputed, with numeric variables filled using the median and categorical variables replaced with "None." Factor variables were converted into a suitable format using one-hot encoding. Finally, numeric features were normalized using the scale () function to ensure consistent scaling across variables.
- **Exploratory Data Analysis (EDA)**
  To identify influential features, a correlation matrix was generated to examine the relationships between variables. Additionally, scatter plots were created to visualize the relationship between auction price and key performance metrics such as runs and wickets. To further explore the impact of categorical variables, box plots were used to analyze the distribution of auction prices across different playing roles. These visualizations are illustrated in the **Visualization** section (Figures 2 to 6).
- **Feature Engineering**
  A new feature, **Is_Bowler**, was derived by combining information from the player's primary role and their wicket count to better capture bowling contributions. Additionally, all categorical columns were factorized to ensure compatibility with modeling algorithms that require numerical input.
- **Model Building / Techniques Used**
  Linear Regression: Linear Regression is a basic predictive modeling technique used to establish a relationship between a dependent variable (in this case, Auction Price) and one or more independent variables (like runs, strike rate, wickets). It assumes a straight-line relationship and is easy to interpret, making it a good starting point for prediction tasks.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n + \varepsilon \quad (1)$$

Where:

y: Target variable (Auction Price)
$x_1, x_2, \ldots, x_n$: Independent variables (e.g., runs, wickets)
β_0: Intercept
$\beta_1, \ldots, \beta_n$: Coefficients for each predictor
ε: Error term

Random Forest: Random Forest is an ensemble machine learning method that builds multiple decision trees and combines their outputs to improve prediction accuracy and control overfitting. It handles both linear and non-linear relationships well and is robust against noise and missing values, making it suitable for complex datasets like player performance data.

$$\hat{y} = (1 / M) \sum_{i=1}^{M} \hat{y}_i \quad (2)$$

Where:

$\hat{y}$: Final predicted auction price
M: Number of decision trees in the forest
$\hat{y}_i$: Prediction from the i-th decision tree

Gradient Boosting Machine (GBM): Gradient Boosting Machine is another ensemble method that builds models sequentially, where each new model corrects the errors of the previous ones. It often results in highly accurate predictions and is particularly powerful for capturing complex patterns in data, though it may require careful tuning to avoid overfitting.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (3)$$

Where:

$F_m(x)$: Prediction after m boosting rounds
$F_{m-1}(x)$: Prediction from previous model iteration
γ_m: Learning rate (shrinkage factor)
$h_m(x)$: Weak learner fitted to residuals

- **Model Evaluation**
  The performance metrics used are RMSE and MAE. RMSE calculates the average of squared differences between predicted and actual values, giving more weight to larger errors. MAE, in contrast, measures the average absolute difference, offering a balanced view of overall prediction accuracy.
- **Technologies and Tools**

This study utilizes R for all analytical stages, including preprocessing, visualization, modeling, and evaluation. Key libraries used include:

- **dplyr**: For data manipulation, filtering, and transformation.
- **ggplot2**: For creating informative visualizations such as histograms, scatter plots, and boxplots.
- **caret**: To streamline the model-building process, including data partitioning, feature engineering, and performance evaluation.
- **randomForest**: To build ensemble-based Random Forest regression models.
- **gbm**: For implementing Gradient Boosting Machines to enhance prediction accuracy.
- **corrplot**: To visualize correlations between numeric variables for deeper feature insights.

This combination of tools supports a robust and comprehensive approach to predicting IPL auction prices,

addressing previously identified gaps such as the lack of auction-specific prediction models, limited use of advanced data preprocessing, and underutilization of ensemble methods in this context. The complete flow of the procedure has been completed, as shown in Figure 1.

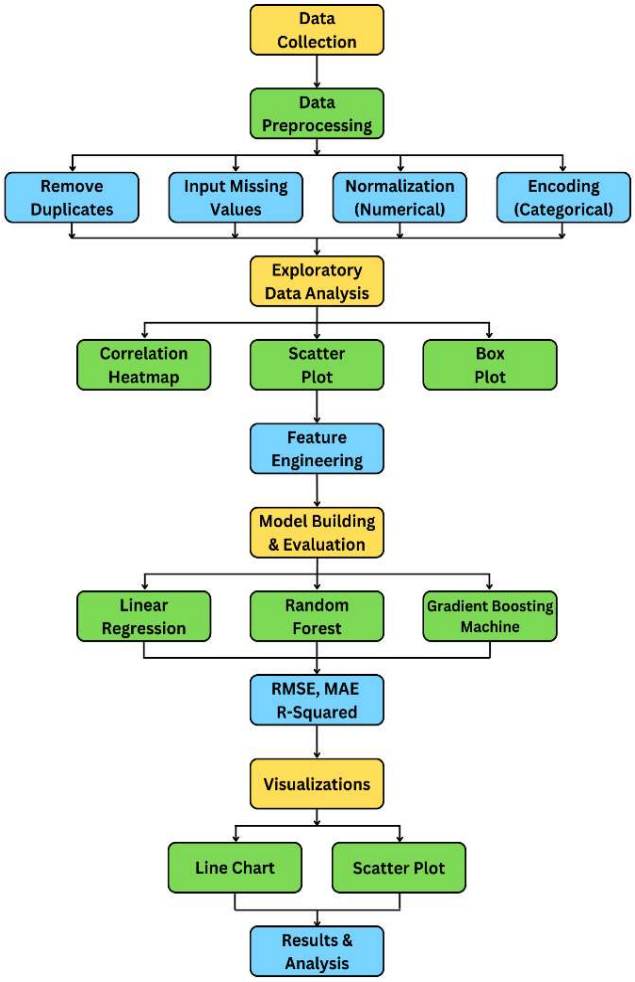

Figure 1: Flow of the prototype

RESULT

The analysis revealed that key performance indicators such as runs scored, strike rate, and wickets taken have a significant impact on a player's auction price in the Indian Premier League (IPL). Among the machine learning models tested, ensemble methods—specifically Random Forest and Gradient Boosting Machine (GBM)—demonstrated superior predictive accuracy compared to traditional linear regression models. These advanced techniques were more effective in capturing complex, non-linear relationships within the data, thereby improving the reliability of the predictions. Overall, Random Forest emerged as the most accurate model for estimating player auction prices, offering strong performance across various evaluation metrics. As shown in Table 1, Random Forest emerged as the most accurate model for estimating player auction prices, offering strong performance across various evaluation metrics.

Table 1: Model Performance Comparison

Model	RMSE	MAE	R-Squared
Linear Regression	5.84	4.28	0.23
Random Forest	4.53	3.61	0.53
Gradient Boosting	4.92	3.78	0.45

VISUALIZATIONS

The histogram shows the distribution of player auction prices, highlighting the frequency of various price ranges and indicating a skewed distribution with most players clustered at lower price points (as shown in figure 2).

Figure 2: Distribution of Auction Prices

The scatter plot illustrating the relationship between base price and final auction price, indicating whether higher base prices generally lead to higher auction outcomes (as shown in figure 3).

Figure 3: Auction Price vs. Base Price

The box plot compares auction prices across different playing roles, revealing variation in valuation depending on the player's primary role (as shown in figure 4).

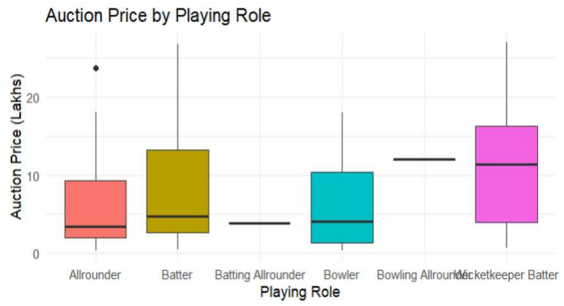

Figure 4: Auction Price by Playing Role

This scatter plot visualizes how auction price correlates with both runs scored and wickets taken, helping to identify whether batting or bowling performance is more strongly tied to market value (as shown in figure 5).

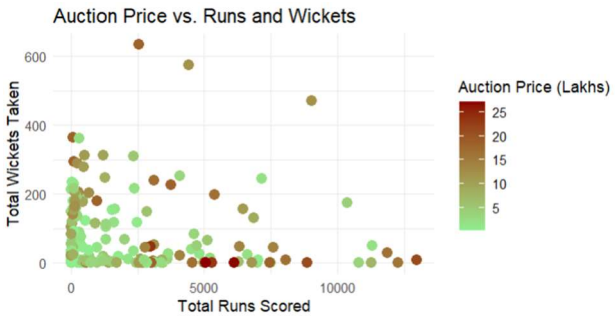

Figure 5: Auction Price vs. Runs and Wickets

The heatmap displaying correlations among numerical variables in the dataset, aiding in identifying which performance metrics have the strongest associations with auction price (as shown in figure 6).

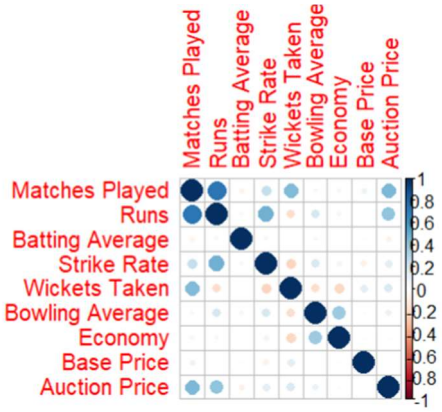

Figure 6: Correlation Heatmap

CONCLUSION

The study successfully carried out comprehensive data preprocessing and cleaning to ensure the quality and reliability of the dataset used for modeling. Multiple machine learning algorithms were developed and rigorously evaluated to identify the most effective predictive approach. A robust prediction system was ultimately built using the Random Forest algorithm, which outperformed other models in accuracy and consistency. Additionally, the research provided valuable visualization insights into trends and patterns in player valuations, enhancing interpretability and aiding in strategic decision-making for IPL franchises.

FUTURE SCOPE

- Include multi-season data and real-time updates.
- Incorporate player popularity and social media metrics.
- Develop a Shiny-based interactive web app for broader usage.

REFERENCES

[1] I. O. Site, "Indian Premier League - IPLT20.com," 25 4 2025. [Online]. Available: https://www.iplt20.com/.

[2] ESPNcricinfo, "IPL Player Statistics," 25 4 2025. [Online]. Available: https://www.espncricinfo.com/.

[3] M. Kuhn, "caret: Classification and Regression Training," [Online]. Available: https://cran.r-project.org/package=caret.

[4] A. a. K. A. Jain, "Machine Learning in Cricket Analytics: A Review," in 2020 International Conference on Computing and Data Science (ICCDS), 2020.

[5] Y. a. S. R. E. Freund, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and System Sciences, pp. 119-139, 1997.

[6] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, New York: Springer-Verlag, 2016.

[7] N. P. Desai, "Assessing Relationship between Auction Price and Performance of Players in Indian Premier League," Journal of Business Analytics and Data Visualization, vol. 6, no. 1, p. 27–37, 2025.

[8] M. F. S. M. G. G. K. R. C. a. B. L. H. R. Prodduturu, "IPL Match Winning Analysis with Player Stats Using Machine Learning," in International Conference on Innovative Approaches in Engineering & Technology (ICIAET-24), 2025.

[9] A. V. V. A. J. A. R. M. M. M. a. P. V. Z. M. Ghayaz, "Performance Evaluation using IPL Performance Impact Model," Journal of Information Systems Research and Practice, vol. 2, no. 5, pp. 2-13, 2024.

[10] V. G. a. V. Bharathi, "Comprehensive Data Analysis and Prediction on Indian Premier League Using Machine Learning Techniques," International Journal of Engineering Techniques, vol. 10, no. 3, 2024.

[11] S. N. P. Y. L. a. F. A. S. Sanjaykumar, "Predicting Team Success in the Indian Premier League Cricket 2024 Season Using Random Forest Analysis," Physical Education Theory and Methodology, vol. 24, no. 2, 2024.

[12] P. R. D. a. A. K. P. Paul, "Classification Model to Predict the Outcome of an IPL Match," in Communications in Computer and Information Science (AGC 2023), Springer, Cham, 2024, pp. 83-111.