A Data-Driven Approach to Identify Future Star Batsmen in the Indian Premier League

Student Name: Kranthi Kumar Yedla
Professor: Mr. Vikas Sahni
Student ID: x2328892
Module: Domain Applications (MSCDAD_C)
National College of Ireland, Dublin

1.Abstract:

Data analytics plays a crucial role in modern cricket. The Principles of Data Analytics and Machine learning technique are applied to this project to identify and classify the young elite batsmen in T20 cricket using detailed IPL match data. Features like strike rate, boundary %, and dot ball % are engineered to capture performance in powerplay, middle overs, and during death overs to determine the best future batting talents. Finally, a Random Forest classifier is used to classify players into top order, middle order, and finisher roles with. This project enables data driven decision making in team selection and scouting, which helps in creating a strong business value.

2.Introduction:

The fusion of data analytics and sports has greatly impacted the cricket team's point of view for evaluating and nurturing talent in the era of data driven decision making. For example, in T20 cricket where the game moves at such a pace, analytics give you a measurable edge and enhances decision making. The Indian Premier League (IPL) is the most competitive cricket tournament. IPL's own rich dataset can yield hidden gems. In this project, predictive analytics is used to select elite batsmen among the young players who debuted after 2020 and classifying them into different roles (top order, middle order or finisher). This project empowers franchises to optimize scouting, investment, and long-term planning in the context of the phase wise (power play, middle and death overs) performance metrics in a match rather than relying on overall statistics.

3.Project Goals:

The purpose of this project is to develop a Machine Learning model by using the principles of Data Analytics to predict elite young batsmen in the Indian Premier League (IPL) from historical match performance data. The term is defined that of cricket specific performance thresholds.

- Strike Rate > 130
- Fewer than 60 IPL matches played
- Scored more than 400 runs.
- Must have debuted on or after the 2020 season of the IPL.

To achieve this, the project:

- It preprocesses and aggregates real world ball by ball and match level data.
- Relevant features like strike rate, runs per match and debut year are the features considered.
- Trained and evaluated a Random Forest Classifier on players data to determine if the predicted elite batsmen are top order, middle order or finisher.

- It is used for visualizing the performance insights of key players to help in scouting decisions.

4. Business Value: This project delivers a solid Business Value in multiple areas which include:

i. Cost Efficiency and Auction Advantage:

This project reveals the underutilized talents, and thus teams are able to sign the underpriced players before they turn into expensive players. With the IPL salary cap, this opens up strategic budget optimization.

ii. Enhanced Team Performance:

The phase-wise strike rates and dot ball percentages give the chance for teams to choose players based on the tactical roles particularly power hitting or death over resilience. It forms a well-balanced winning squad.

iii. Increased Revenue Streams:

The more attention they draw in terms of young players with rising star power, the greater the audience interest can be and the higher ticket sales and engagement with merchandise and digital content. Exotic emerging talent are more likely to align themselves with brands and sponsors.

iv. Strategic Longevity:

This continuity in performance continues to be ensured by consistent identification of new talent as the veteran players retire. It assures the long life of the project and fan loyalty, stakeholder confidence across seasons.

v. Brand Value and Marketability:

Franchises like CSK have amassed their identity through a couple of marquee players such as MS Dhoni. Yet their roster today has fewer players who are brand builders in their prime. One could over-rely on one figure of note, which could diminish the strength of the brand long term.

If teams such as CSK invest in smarter talent scouting and the future stars, they can sustain fan engagement, build new brand ambassadors and make the team more marketable for years to come.

vi. Franchise Global Expansion:

Mumbai Indians have created global business footprints by having strong scouting system by identifying and nurturing local talent into international stars, partnerships, fan bases and media coverage are all positively enhanced.

5.Literature Review and Techniques Used in Each Paper:

1.Applications of Machine Learning in Cricket: A Systematic Review (2021)

- This research reviews twenty-five different ML/AI algorithms used for cricket analytics which encompass SVM, Decision Trees, ANN, KNN, Naive Bayes, Random Forest, and Genetic Algorithms.
- Focus: Highlights predictive performance, team selection, match outcome prediction, and talent identification across various papers.

- Random Forest ensemble models achieve superior predictive performances than individual learners because they provide better generalization results.
- Reference: [1] Applications of ML in Cricket.

2.Cricket Team Selection and Player Analysis Using Data Analytics (2021)

- The methodology applies K-Means Clustering algorithms for both team player segmentation and performance measurement tasks.
- The analysis involved statistical data evaluation together with cluster-based separation of player attributes.
- Limitation: Unsupervised; no prediction or classification capability.
- Reference: [2] Cricket_Team_Selection_and_Player_Analysis

3.The Identification of Game Changers in England Cricket's Developmental Pathway (2019)

- Random Forest machines, Support Vector Machine (SVM) and Multilayer Perceptron together with Naive Bayes, Recursive Feature Elimination (RFE) served as the technique(s) for this analysis.
- The research team applied pattern recognition alongside feature engineering procedures to 93 features for modelling both elite and sub-elite players.
- Random Forest demonstrated successful field deployment because of its high accuracy levels alongside good interpretability capabilities which were validated through external testing.
- Reference: [3] Game Changers – Jones et al., 2019

4.OWA Based Model for Talent Selection in Cricket (2014)

- Technique(s): Ordered Weighted Averaging (OWA) with Fuzzy Linguistic Quantifiers.
- Focus: Multi-criteria decision-making using subjective weights and fuzzy logic.
- The process has limitations because it predicates on human evaluations using predefined test scores yet lacks data-driven prediction capabilities.
- Reference: [4] OWA Based Model

Final Verdict:

According to the Jones et al. [3] study in the Game Changers, Random Forest Classifier was the best model for talent identification in cricket. With Recursive Feature Elimination (RFE) and pattern recognition across 93 engineered features, it was able to perform with an accuracy of 92.9%. This model was more accurate, had better generalization ability, and was interpretable, when one is trying to assess performance of complex data in sports. The Random Forest outperforms other models mentioned in the literature such as Unsupervised K-Means Clustering [2] and OWA based fuzzy logic models [4] as they rely on subjective weights and require subjective inputs. As the most relevant and dependable one for a goal of classifying young IPL batsmen as per their respective performance metrics, its ensemble learning structure prevents overfitting and always guarantees consistent results. Moreover, systematic ML reviews are provided for cricket analytics [1].

6. Dataset Description:

The project is based on the two IPL datasets matches.csv and deliveries.csv. In particular, the matches.csv includes high-level match information, including the season, teams involved, winner, and the venue, and is used to find players' debut year. The deliveries.csv file contains ball by ball details like runs scored, wickets, overs and player actions. Performance analysis of this dataset makes it possible to generate such key features such as strike rate percentage, boundary percentage and dot ball percentage across different game phases. These datasets are used together to identify and classify young elite batsmen using machine learning.

7. Exploratory Data Analysis:

For the understanding of player trends and potentially identifying relevant performance patterns, EDA was carried out. The project has aligned with the focus of looking for the increasing number of young players entering the IPL, which can be proved with a bar chart of debut years:

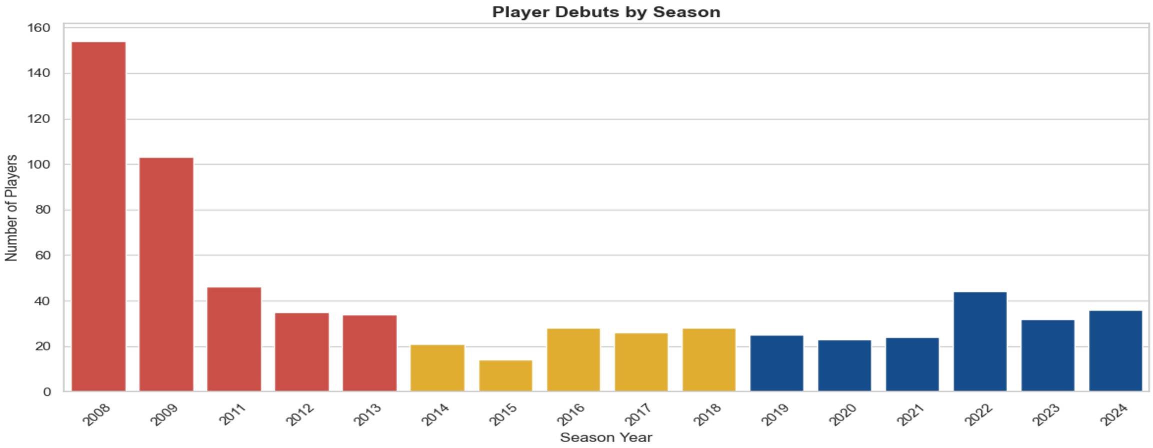


Figure 1

To detect and analyze the difference in the performance between Powerplay, Middle and Death overs, phase wise metrics like strike rate, boundary percentage and dot ball percentage were visualized:

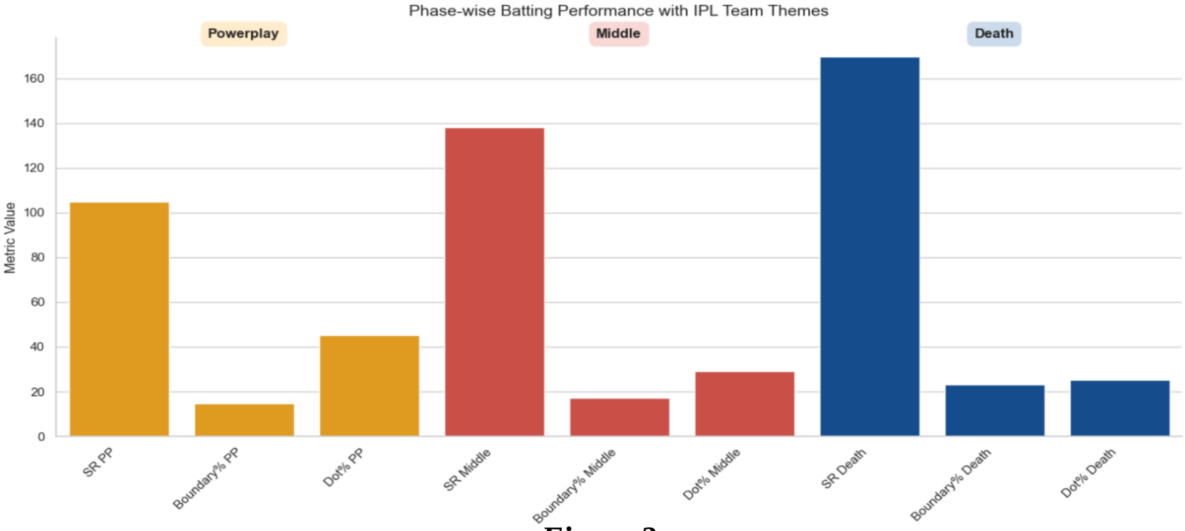


Figure 2

Strategic insights about role-based classification and talent scouting are supported by these visuals. A feature engineering problem was defined to extract batting metrics during Powerplay, Middle and Death overs such as strike rate, boundary percentage and dot ball percentage. Features captured in these phase-wise manner captured situational performance and allowed to accurately classify players to roles. It was critical to this training of a reliable and role-aware machine learning model:

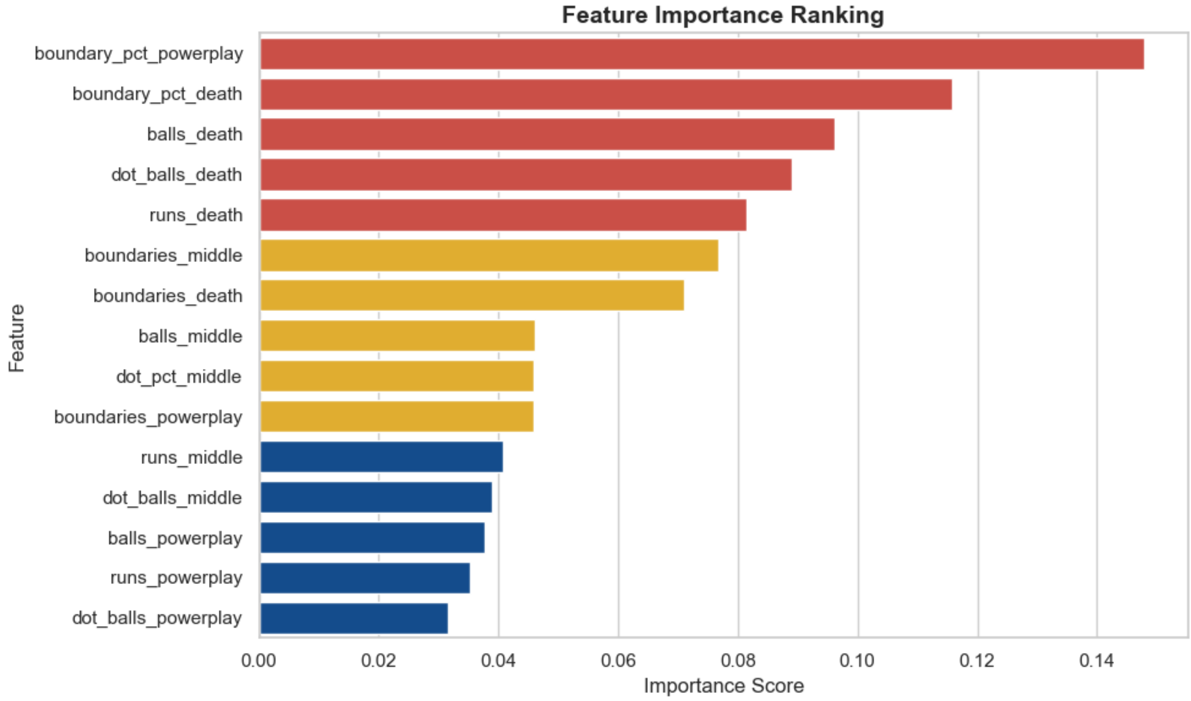


Figure 3

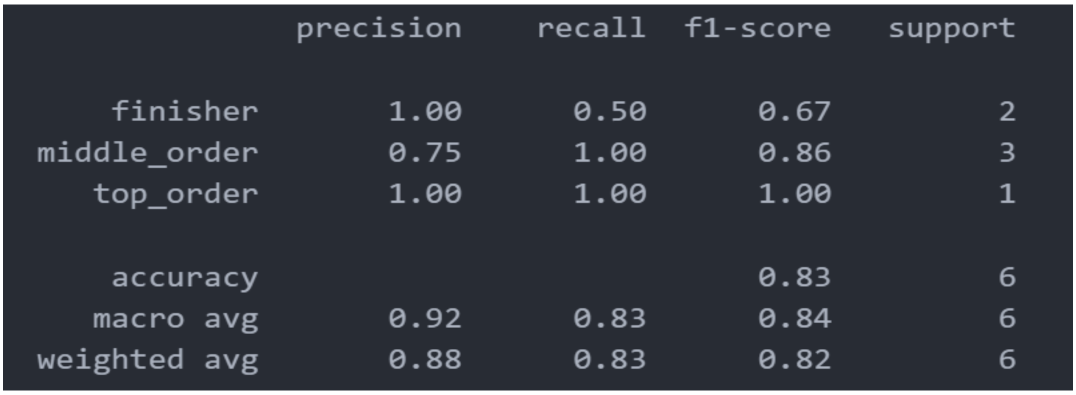
8. Methodology:

To this end, the methodology processes ball by ball IPL data to predict and classify young elite batsmen using machine learning. Then, delivery.csv dataset is merged with matches.csv to get each player's debut year. Built in phase of features i.e. the Powerplay (1-6), Middle (7-15) and Death overs (16-20), strike rates, boundary percentage, dot ball percentage are decided accordingly. Performance thresholds are set for players to be labelled as elite and they are forwarded into their classification roles (i.e., top order, middle order or finisher). A Random Forest Classifier is trained over these features with 80% of the data used for train and other 20% used for test (a stratified 80-20 train test split). Accuracy, classification report and confusion matrix are used to evaluate model performance. We use visualizations for interpreting role distributions, feature importance, and trends in player debuts.

9. Implementation of Machine Learning Technique:

Training the supervised learning algorithm for this model implementation had been done on young elite batsmen to classify them into top order, middle order or finisher roles. There are several options for picking a classifier; one fairly robust, and fairly interpretable, is a Random Forest Classifier. The input features used were phase wise batting metrics such as strike rate, boundary percentage, and dot ball percentage for Powerplay, Middle, and Death overs. Those players with a debut year after 2020 and less than 60 matches were shortlisted. Finally, the dataset was split using an 80-20 stratified train test split, so that the role distribution is preserved. The StandardScaler was used to make features standardized and used 100 estimators and max depth of 6 to train the model. Performance was assessed in terms of accuracy, classification report and a confusion matrix with respect to accuracy of 83%, suggesting strong capability to classify players in roles.

The Figure 4 describes the classification report of the implemented Random Forest Classifier model:



```
              precision    recall  f1-score   support

    finisher       1.00      0.50      0.67         2
middle_order       0.75      1.00      0.86         3
   top_order       1.00      1.00      1.00         1

    accuracy                           0.83         6
   macro avg       0.92      0.83      0.84         6
weighted avg       0.88      0.83      0.82         6
```

Figure 4

The model has classified the players into the following roles:

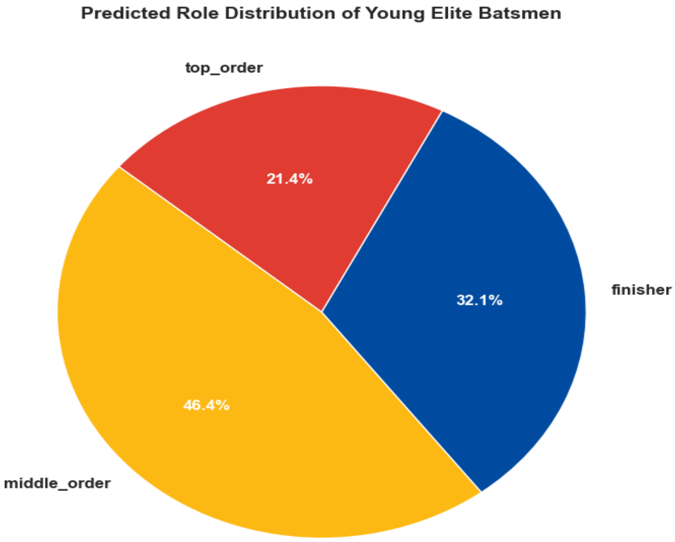
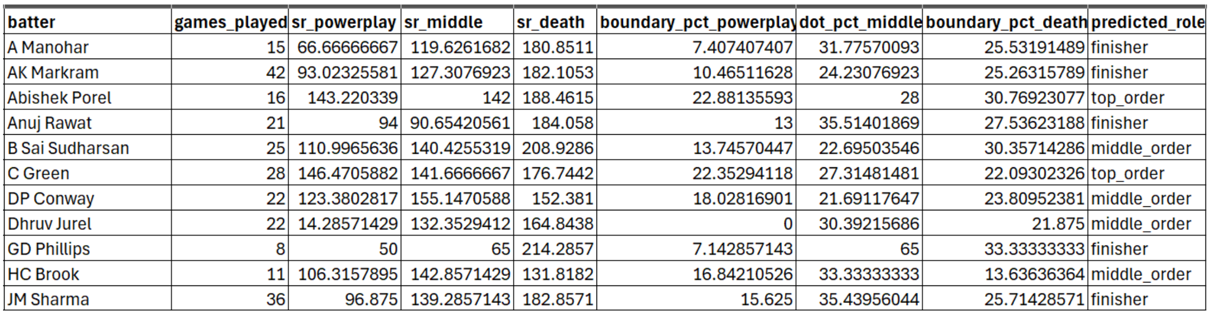


Figure 5

List of Identified Future Elite Batsmen:



batter	games_played	sr_powerplay	sr_middle	sr_death	boundary_pct_powerplay	dot_pct_middle	boundary_pct_death	predicted_role
A Manohar	15	66.66666667	119.6261682	180.8511	7.407407407	31.77570093	25.53191489	finisher
AK Markram	42	93.02325581	127.3076923	182.1053	10.46511628	24.23076923	25.26315789	finisher
Abishek Porel	16	143.220339	142	188.4615	22.88135593	28	30.76923077	top_order
Anuj Rawat	21	94	90.65420561	184.058	13	35.51401869	27.53623188	finisher
B Sai Sudharsan	25	110.9965636	140.4255319	208.9286	13.74570447	22.69503546	30.35714286	middle_order
C Green	28	146.4705882	141.6666667	176.7442	22.35294118	27.31481481	22.09302326	top_order
DP Conway	22	123.3802817	155.1470588	152.381	18.02816901	21.69117647	23.80952381	middle_order
Dhruv Jurel	22	14.28571429	132.3529412	164.8438	0	30.39215686	21.875	middle_order
GD Phillips	8	50	65	214.2857	7.142857143	65	33.33333333	finisher
HC Brook	11	106.3157895	142.8571429	131.8182	16.84210526	33.33333333	13.63636364	middle_order
JM Sharma	36	96.875	139.2857143	182.8571	15.625	35.43956044	25.71428571	finisher

Figure 6

10. Ethical Considerations:

The data used for this project is publicly available IPL data and no such personal information is exposed and preserved the data privacy. For one, machine learning helps identify talent, and predictive analytics provides pertinent insights too, but it is still questionable as to why one would rely entirely on machine learning methods when there is a simple data analytics approach that could be easily adopted and capitalized upon. It may miss late bloomers or new player profiles not measured by standard metrics. The model can not replace, but support expert judgment. For that, we need transparency of the model limits and responsible communication of results in the case of ethical deployment. This helps in avoiding stereotypes as well as ensuring fairness with human oversight, to avoid biased decision making and to democratise talent assessment in cricket.

11. Conclusion and Future Work:

The machine learning framework for identifying elite young batsmen in the Indian Premier League was demonstrated in this project. The model was able to successfully determine high potential players from the real IPL match data when the data is fed to a Random Forest classifier. The model was quite predictive and identified players like Jitesh Sharma, Tim David, and Rinku Singh as the emerging ones. It powered interpretable knowledge for talent scouts and decision makers to act on actionable insights. However, the model provides good support for modelling talent identification while only itself excludes batting performance. For example, future iterations could include other dimensions, such as match context, opposition strength, venue effects, psychological assessments in addition to these. There are potential next steps to integrating live data streams and deploying the solution as an interactive scouting dashboard. In general, this is a demonstration of the practical value of machine learning and how it can be used in the sports analytics domain to enable IPL stakeholder decisions that are informed and based on evidence.

12. References

[1] Shaikh, T., Khandare, R., & Bhosale, M. (2021). *Applications of Machine Learning in Cricket: A Systematic Review*. International Research Journal of Engineering and Technology (IRJET). Link

[2] Rai, R., & Kumar, R. (2021). *Cricket Team Selection and Player Analysis Using Data Analytics*. International Journal of Advanced Research in Computer and Communication Engineering. Link

[3] Jones, B. D., Lawrence, G. P., & Hardy, L. (2019). *The Identification of 'Game Changers' in England Cricket's Developmental Pathway: A Mixed Methods Approach*. Journal of Expertise. Link

[4] Ahamad, M., Dey, L., & Singh, A. (2014). *OWA Based Model for Talent Selection in Cricket*. International Journal of Computer Applications. Link

Dataset: https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020?select=deliveries.csv