Capstone Project Report: IPL Data Analysis

Akhil Inaganti, Sachin Kashyap, Sai Venkata Krishna Reddy Boyapati, Sumanth Reddy Gajjala, Surya Narasimha Manikanta Pavankumar Akula Jaya

Introduction

Cricket is not just a sport in India; it's a religion, with millions of followers and an immense impact on the country's culture and economy. The Indian Premier League (IPL), a professional Twenty20 cricket league, is among the most celebrated sports events globally. It brings together the world's best cricketing talent to compete in an annual tournament that captivates audiences across the globe.

Our project's goal is to analyze IPL match data to identify the best players suited for a new tournament. The aim is to help brands select these standout performers as brand ambassadors, leveraging their popularity and influence. The motivation behind this project stems from cricket's significant role in marketing and advertising in India. Cricket stars are not just athletes; they're icons that wield considerable influence over public opinion and consumer behavior. By identifying the most promising and influential players, brands can harness this influence to enhance their market presence and connect with millions of cricket fans.

Data Preprocessing

The primary task was to convert JSON files into a CSV format and then combine all the match data into a single, comprehensive dataset. This process involved reading individual JSON files, each representing data from an IPL match, extracting relevant details such as match information, player performances, and ball-by-ball actions. The extracted information was then structured and aggregated into a unified CSV file, providing a consolidated view of all matches. This step was crucial for facilitating further analysis, enabling a more straightforward exploration of the data and application of machine learning models to achieve the project's goal of identifying the best players for potential brand ambassador roles.

In next phase focused on cricket metrics, you conducted an in-depth analysis to calculate key performance indicators for both batsmen and bowlers. Here's a summary of the steps you took:

Batsmen Metrics:

- **Runs Scored and Balls Faced:** Aggregated the total runs scored and balls faced by each batsman.
- **Dismissals:** Counted how many times each batsman got out to calculate their batting average.
- **Batting Average:** Calculated as total runs scored divided by the number of times dismissed.
- **Strike Rate:** Calculated as runs scored per 100 balls faced.

Bowlers Metrics:

- **Deliveries and Runs Conceded:** Aggregated the total deliveries bowled and runs conceded by each bowler.
- **Wickets Taken:** Counted the number of wickets taken by each bowler, excluding runouts.
- **Bowling Average:** Calculated as total runs conceded divided by wickets taken.
- **Economy Rate:** Calculated as runs conceded per over.
- **Bowling Strike Rate:** Calculated as the number of balls bowled per wicket taken.

We also handled missing values by filling in zeros for bowlers with no wickets and batsmen with no dismissals to ensure the calculations remained accurate. This comprehensive analysis provides a solid foundation for evaluating player performance, enabling the identification of top performers for potential selection as brand ambassadors.

Exploratory Data Analysis (EDA)

The basic statistics of the dataset reveal insights into both batting and bowling metrics for a set of 694 players:

Batting Metrics:

- **Runs Scored:** Players scored between 0 to 7273 runs, with an average of 440 runs.
- **Balls Faced:** Ranges from 0 to 5739 with an average of 351 balls faced.

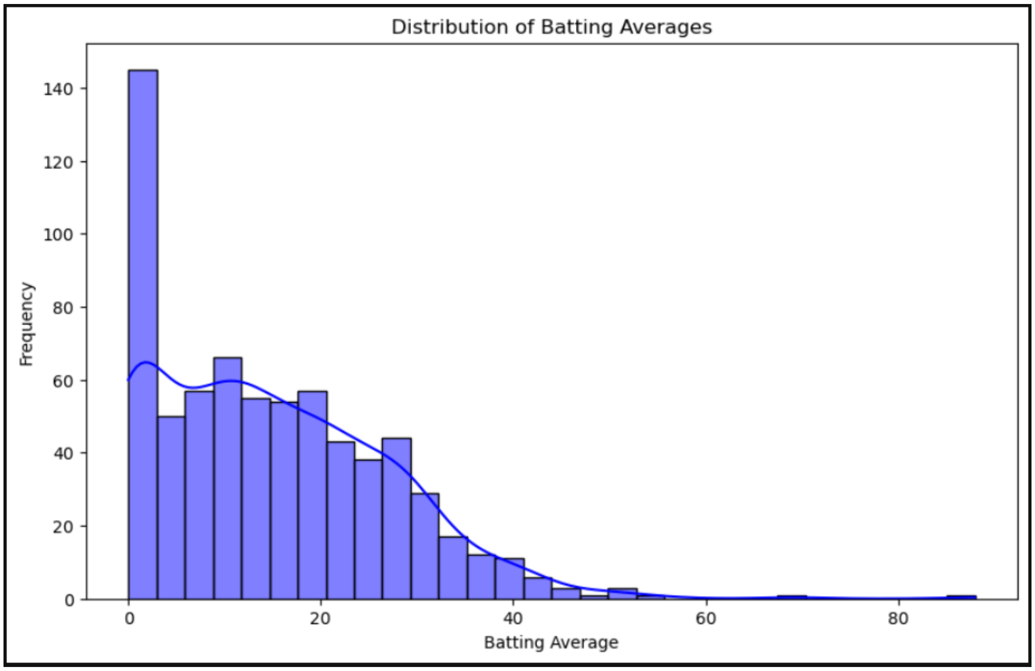
- **Dismissals:** Varies from 0 to 208, with an average of approximately 17 dismissals per player.
- **Batting Average:** Has a wide range from 0 to 88, with a mean batting average of about 14.83.
- **Strike Rate:** Shows a significant variation from 0 to 250, with a mean strike rate of 93.27.

Bowling Metrics:

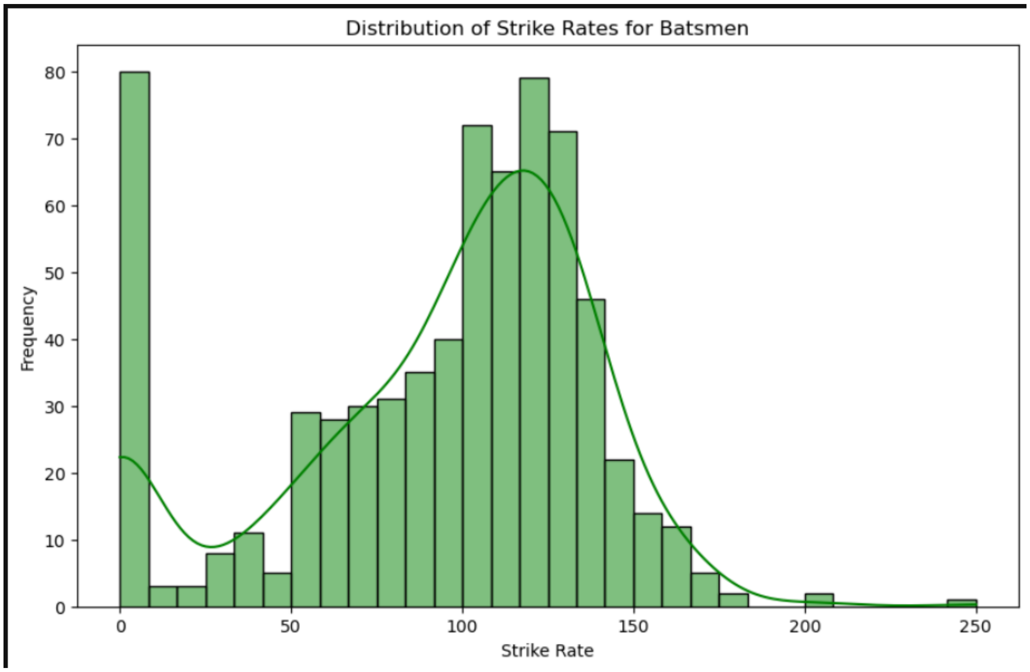
- **Balls Bowled:** Players bowled between 0 to 4333 deliveries, with an average of 351 deliveries.
- **Runs Conceded:** The number of runs conceded ranges from 0 to 4739, with an average of 440 runs.
- **Wickets:** The number of wickets taken ranges from 0 to 187, with an average of about 16 wickets per player.
- **Bowling Average:** Ranges significantly from 0 to 132, with an average bowling average of 20.77.
- **Economy Rate:** Varies widely from 0 to 36, with a mean economy rate of 6.05.
- **Bowling Strike Rate:** Shows a broad range from 0 to 88, with an average strike rate of 15.56.

Insights & Patterns:

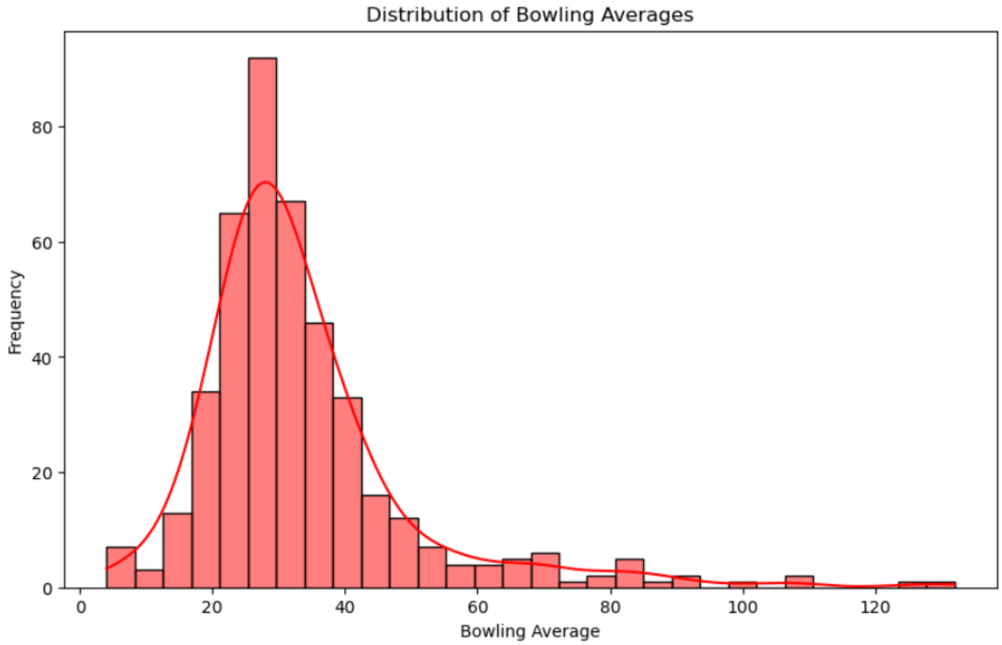
- There's a wide range in the performance metrics of both batsmen and bowlers, indicating a diverse set of players in the dataset.
- The high standard deviations in runs scored, balls faced, and wickets indicate significant variability in player performances.
- Some players have exceptionally high or low values in certain metrics (like strike rate and economy rate), which might indicate outliers or particularly noteworthy performances.



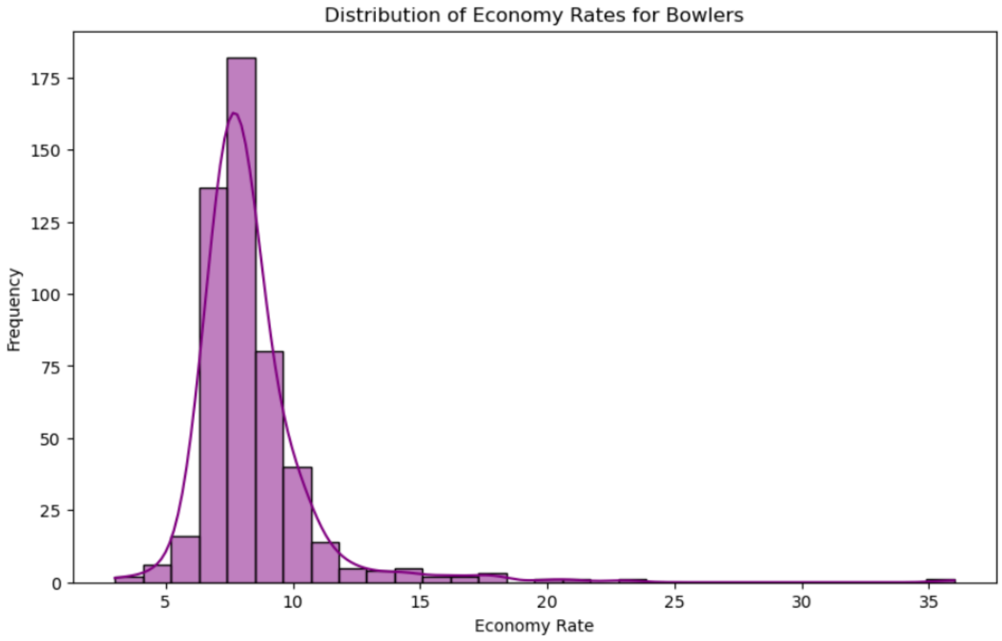
The distribution is skewed to the right, indicating that while most players have a lower batting average, a few exceptional players achieve very high averages.



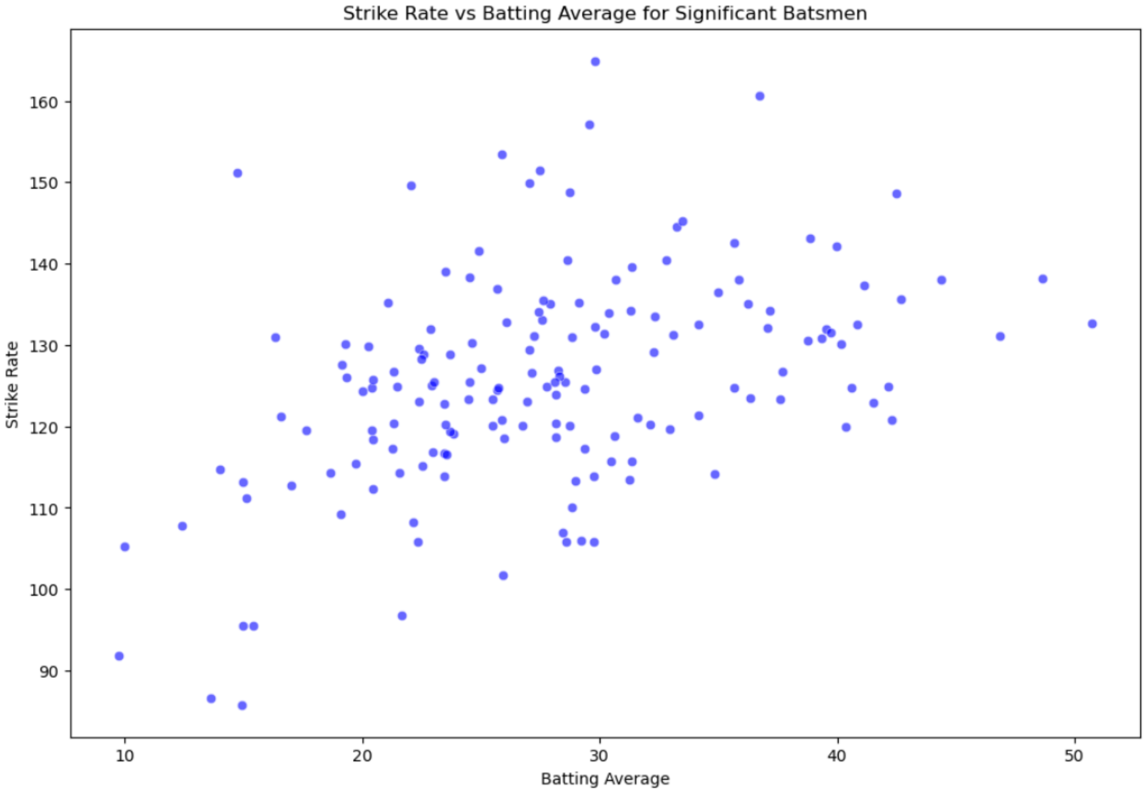
This distribution is more varied, with peaks around lower and higher values, suggesting the presence of both conservative players and aggressive hitters.



The bowling average distribution is also right-skewed, similar to batting averages, showing that most bowlers have higher averages with a few outstanding performers with lower averages, indicating better bowling performance.



The economy rate distribution shows a peak around the 6-8 runs per over range, which is typical for T20 cricket. Lower values indicate more economical bowlers, while higher values are less desirable.



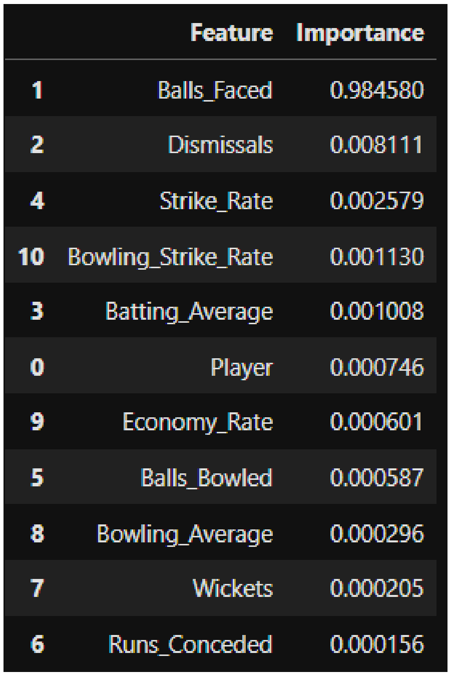
The scatter plot reveals a variety of batting styles. A cluster of players with high batting averages and moderate strike rates suggests consistent performers. In contrast, players with high strike rates but lower averages might be seen as aggressive hitters, potentially useful in specific match situations.

These analyses offer valuable perspectives for team selection, strategy formulation, and identifying players with potential for standout performances in specific match situations.

Machine Learning Models

For the machine learning aspect of your project, we are still in the exploratory data analysis (EDA) phase, aiming to understand which features are most important for predicting player performance. Feature importance is a crucial step in building effective machine learning models as it helps in selecting the most relevant features that contribute to the prediction.

Feature Importance by using Random Forest model.



	Feature	Importance
1	Balls_Faced	0.984580
2	Dismissals	0.008111
4	Strike_Rate	0.002579
10	Bowling_Strike_Rate	0.001130
3	Batting_Average	0.001008
0	Player	0.000746
9	Economy_Rate	0.000601
5	Balls_Bowled	0.000587
8	Bowling_Average	0.000296
7	Wickets	0.000205
6	Runs_Conceded	0.000156

Removing features like 'Batting Average' and 'Strike Rate' due to their high correlation with other features is a wise step in enhancing your machine learning model's performance.

Through rigorous analysis and machine learning modeling, our project has effectively utilized IPL match data to identify promising players who could serve as brand ambassadors. Our suite of machine learning models, including Linear Regression, Lasso, Ridge, Decision Tree, Random Forest, and Gradient Boosting, were applied to predict player performance metrics accurately.

Model	R2 Score	MSE	MAE	RMSE
Linear Regression	0.870731	55.716686	5.293132	7.464361
Lasso	0.865236	58.085043	5.313315	7.621354
Ridge	0.870759	55.704864	5.291518	7.463569
Decision Tree Regressor	0.742072	111.170781	7.074811	10.543756
RandomForest Regressor	0.855995	62.068132	5.343269	7.878333
Gradient Boosting Regressor	0.874423	54.125370	5.095255	7.356995

Conclusion

R2 Score: 0.874423 — This is the highest among all the regression models, suggesting that it's the most capable of capturing the variance in the dataset.

MSE: 54.125370 and RMSE: 7.356995 — Both metrics are the lowest among the regression models, indicating that this model has the best predictive accuracy with the least amount of error.

MAE: 5.095255 — Also the lowest, which means on average, the predictions are very close to the actual data points.

Opting for the Gradient Boosting Regressor is top performance in our analysis. This model has demonstrated the best balance between complexity and accuracy, with the highest R2 Score and the lowest error metrics among the models.

Recommendations

Based on our findings, we recommend the following for leveraging IPL player data in brand endorsement strategies:

Focus on Top Performers: Brands should consider players who consistently show high performance metrics as identified by the Gradient Boosting Regressor. These players are likely to have a substantial impact and a broad appeal among fans, enhancing brand visibility and engagement.

Diverse Player Profiles: Incorporate a mix of player profiles, including both consistent performers and potential high-impact players. This strategy allows brands to cover various aspects of the game and appeal to different audience segments.

Data-Driven Decisions: Regularly update the analytical models with new data each season to keep the brand ambassador selection relevant and effective. This approach ensures that the endorsements reflect current performances and market trends.

Contributions

Sai Venkata Krishna Reddy Boyapati- Worked on the gathering the data from twitter.

Sumanth Reddy Gajjala- Worked on the converting the ball-to-ball data to match data and machine learning models.

Akhil Inaganti- worked on random forest model.

Sachin Kashyap- worked on getting the data from cricsheet.org and converting the json file to csv.

Surya Narasimha Manikanta Pavankumar Akula Jaya – Worked on machine learning model using impact runs.

Reference

Data - https://cricsheet.org/

IPL - https://www.iplt20.com/

Databox - https://databox.com/data-analysis-report