Visualizing Key Performers in Cricket

Project Report

Group No 2

Sumanth Reddy Gajjala 0816129

Surya Narasimha Manikanta PavanKumar Akula Jaya 0811656

Sachin Kashyap 0832574

Sai Venkata Krishna Reddy Boyapati 0813504

Akhil Inaganti 0811914

Abstract

In the vast field of sports analytics and marketing, we aim to utilize the influence of cricket players for brand promotion. This study involves identifying optimal cricket players for brand endorsements by evaluating their current influence and predicting future popularity through Twitter analysis. We employ advanced neural network models on textual data extracted from Twitter to gain insights into players' social media presence.

By treating Twitter as a vast source of information, we utilize neural networks—a set of computer algorithms inspired by the human brain—to comprehend sentiments expressed in tweets about players. This complex yet fascinating process helps us discern whether people are expressing positive or negative opinions about the players.

Our objective is to assist businesses in selecting cricket players with sustained popularity and broad public appeal. This research serves as a guide for companies, enabling them to make informed choices in partnering with players who resonate well with the public, ensuring the success of brand collaborations in the ever-evolving landscape of cricket and advertising.

In addition to evaluating cricket players' popularity on Twitter, we will also consider their past performance scores to determine the best players. By analyzing their previous scores, we aim to identify players who not only have a strong social media presence but also demonstrate exceptional on-field capabilities. This dual approach allows us to comprehensively assess and select the best-suited cricket players for brand endorsements, considering both their online popularity and past sporting achievements.

Introduction

Background

Cricket is a popular bat-and-ball sport played between two teams. Each team takes turns batting and bowling, with the aim of scoring more runs than the opposition. The batting team tries to hit the ball and run between wickets, while the bowling team attempts to dismiss the batsmen and restrict runs. Cricket is known for its diverse formats, including Test matches, One Day Internationals (ODIs), and Twenty20 (T20) matches. It has a global following, with passionate fans, iconic players, and a rich history, especially in countries like India, England, Australia, and beyond.

In India, cricket is like a language everyone speaks, connecting people from different places and backgrounds. It's not just a game; it's something that brings us all together. Our research comes from the understanding that cricket is a big part of everyone's heart in India.

Motivation

Our motivation is rooted in the fact that cricket is more than just a sport here; it's a feeling we all share. Whether you're in a city or a small village, everyone loves cricket. The excitement in the stadiums, the talks during matches, and the joy after a win create a special experience that everyone can relate to. Driven by this shared love for cricket, we want to explore how it connects with promoting brands.

Objectives

1. Identifying Influential Players

The primary objective is to identify and rank cricket players based on their current influence, considering factors such as social media engagement, fan interactions, and online popularity.

2. Predicting Future Popularity

We aim to predict the future popularity of players by analyzing Twitter data using advanced neural network models. This involves understanding sentiment, engagement, and trending topics related to each player.

3. Incorporating Performance Scores

In addition to social media analysis, we will integrate players' past performance scores to provide a comprehensive evaluation. This dual approach ensures a balanced assessment, considering both off-field influence and on-field prowess.

4. Guiding Brand Endorsements

The ultimate goal is to provide businesses with a guide for selecting cricket players for brand endorsements. This guide will consider a holistic view, encompassing social media popularity and past performance, to maximize the impact of brand collaborations.

Methodology

Data Sources
Our primary data sources for this research include reputable cricket databases and platforms such as ESPNCricinfo and Cricket Australia. Additionally, we utilized the https://cricsheet.org/ website, where we obtained detailed ball-by-ball data for each IPL match in JSON format.

Data Retrieval
Utilized API endpoints from ESPNCricinfo, Cricket Australia, and https://cricsheet.org/ to gather structured cricket data, covering player statistics, match histories, and detailed ball-by-ball information for IPL matches.

Data Preprocessing
Employed Pandas and NumPy for cleaning, preprocessing, and transforming raw data into a structured format. This involved handling missing values, normalizing data, and encoding categorical variables. Utilized .NET to convert the JSON files obtained from https://cricsheet.org/ into CSV format.

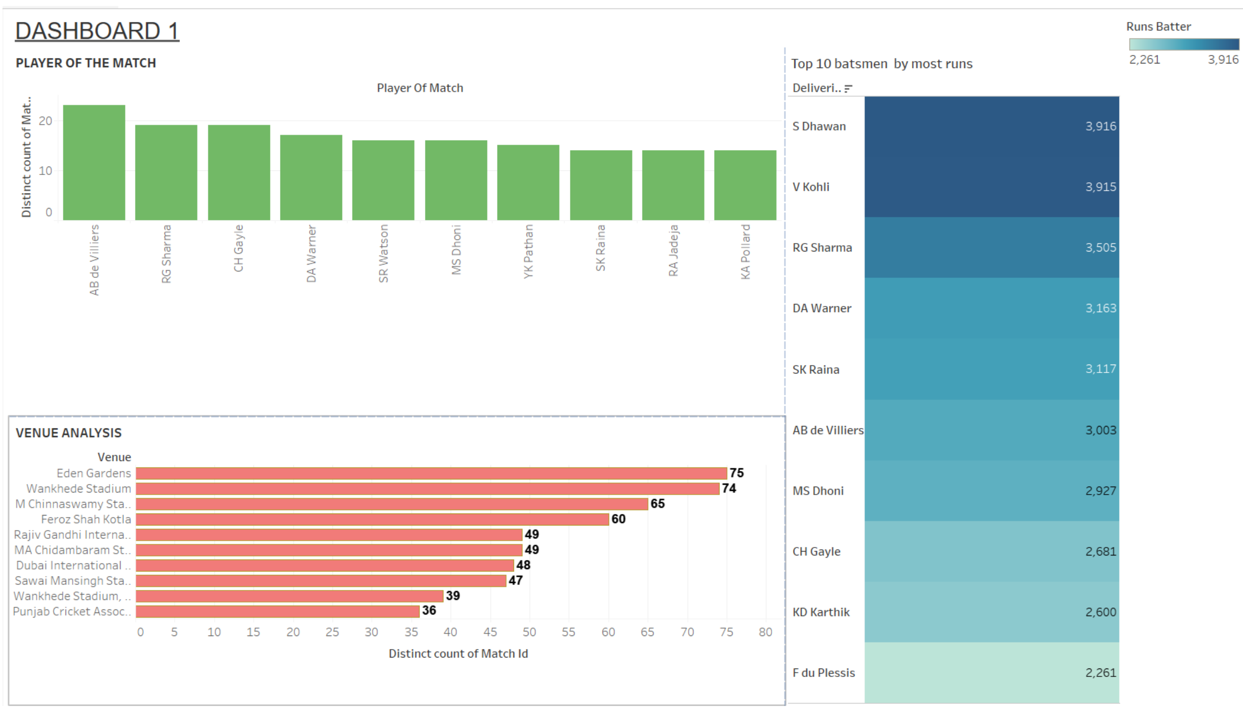
Exploratory Data Analysis (EDA)
Conducted an in-depth exploration of the data, using visualizations and statistical analysis to glean insights into player and team performance, identify outliers, and understand factors influencing team rankings.

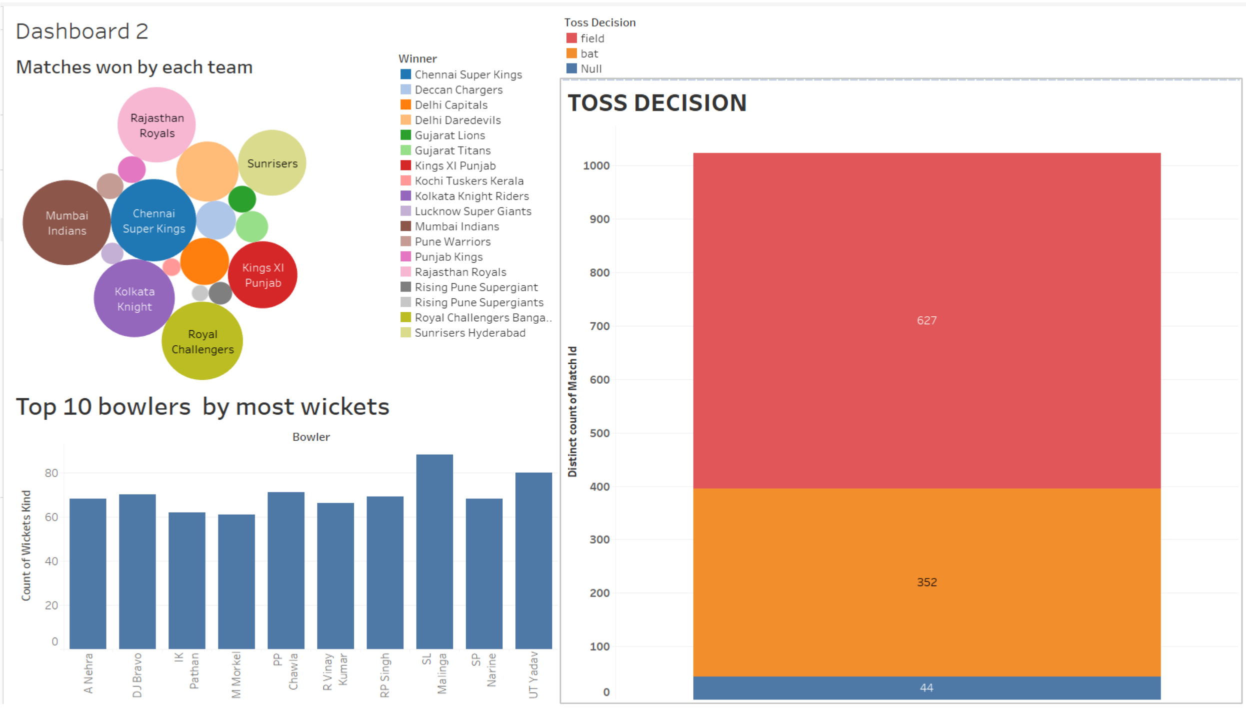
Visualization

Utilized Matplotlib, Seaborn, and Tableau for insightful visualizations, including interactive dashboards.

Our research began with understanding and integrating cricket datasets. Preliminary analysis explored data quality and preprocessing techniques. Twitter data analysis for player popularity is underway, providing initial insights. .NET facilitated seamless conversion of detailed ball-by-ball data into CSV. Visualizations were created in Tableau, enhancing the presentation of key findings and insights. The integration of multiple tools enriches the depth and breadth of our analysis.

Data Analysis and Visualizations





Insights

1. Team Performance:

Consistent Excellence: Mumbai Indians, Chennai Super Kings, and Royal Challengers Bangalore have maintained a remarkable and consistent performance, emerging as top-performing teams with the most victories.

2. Player Recognition:

AB de Villiers and Rohit Sharma: These players have consistently shone with more "Player of the Match" awards, highlighting their significant individual contributions and match-defining performances.

3. Batting Prowess:

Dhawan and Kohli's Run Dominance: Shikhar Dhawan and Virat Kohli's exceptional run-scoring abilities make them key players. Their consistent performances contribute significantly to their team's success.

Toss Strategy:

4. Strategic Decision-Making:

The prevalent trend of choosing to field first in toss decisions suggests a strategic approach by teams. Factors such as pitch conditions, weather, and team strengths likely influence this decision-making.

These insights offer a glimpse into the dynamics of IPL matches, showcasing the dominance of certain teams, the individual brilliance of players, and strategic decisions influencing match outcomes. Understanding these patterns provides valuable knowledge for future match predictions and team strategies.

Interpretations

Team Dominance: The dominance of Mumbai Indians, Chennai Super Kings, and Royal Challengers Bangalore indicates a consistent level of excellence, suggesting well-established team strategies, leadership, and player performances.

Player Excellence: AB de Villiers and Rohit Sharma's frequent "Player of the Match" awards underline their impact on match outcomes. Shikhar Dhawan and Virat Kohli's consistent run-scoring highlights their crucial role as batting mainstays.

Toss Strategies: The common trend of choosing to field first suggests teams might prefer chasing targets, emphasizing adaptability to pitch and weather conditions.

Limitations

Data Quality: The quality of insights heavily relies on the accuracy and completeness of the data. Any inaccuracies or missing information may impact the robustness of conclusions.

Recommendations and Future Work

1. Enhanced Data Collection

Continuously refine data collection methods to ensure comprehensive coverage, considering multiple seasons and diverse datasets.

2. Feature Engineering

Develop novel features capturing nuances of batting, bowling performance, team dynamics, and historical trends. Apply domain knowledge for meaningful metric creation.

3. Machine Learning Models

Implement advanced algorithms (ensemble methods, neural networks, gradient boosting) using Scikit-Learn and TensorFlow. Predicted team performance considering batting averages, bowling strike rates, and match locations.

4. Model Evaluation

Employe cross-validation, grid search, and metrics (MAE, RMSE) to optimize model hyperparameters and assess performance.

5. Twitter Analysis

Finding insights from Twitter sentiment analysis indicate a positive reception for certain players, forming a foundation for predicting future player popularity.

Appendix

EDA

```
ipl_data.info()
```

```
#   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   city                 120350 non-null  object
 1   dates                126855 non-null  object
 2   event_match_number   119104 non-null  float64
 3   outcome_by_runs      35120 non-null   float64
 4   winner               126706 non-null  object
 5   overs                126855 non-null  int64
 6   player_of_match      124804 non-null  object
 7   season               71940 non-null   object
 8   toss_decision        121499 non-null  object
 9   toss_winner          76405 non-null   object
 10  venue                126855 non-null  object
 11  innings_team         126855 non-null  object
 12  overs_over           20280 non-null   float64
 13  deliveries_batter    126001 non-null  object
 14  bowler               126001 non-null  object
 15  non_striker          126001 non-null  object
 16  runs_batter          126001 non-null  float64
 17  extras               126001 non-null  float64
 18  total                126001 non-null  float64
 19  extras_wides         3753 non-null    float64
 20  legbyes              1637 non-null    float64
 21  noballs              695 non-null     float64
 22  wickets_kind         3866 non-null    object
 23  player_out           3866 non-null    object
 24  fielders_name        4840 non-null    object
 25  powerplays_from      126855 non-null  float64
 26  to                   126855 non-null  float64
 27  type                 121766 non-null  object
 28  innings_2_team       123687 non-null  object
```

```
29  overs_over.1          19128 non-null   float64
30  deliveries_batter.1   117445 non-null  object
31  bowler.1              117445 non-null  object
32  non_striker.1         117445 non-null  object
33  runs_batter.1         117445 non-null  float64
34  extras.1              117445 non-null  float64
35  total.1               117445 non-null  float64
36  extras_wides.1        1585 non-null    float64
37  noballs.1             137 non-null     float64
38  legbyes.1             558 non-null     float64
39  wickets_kind.1        3521 non-null    object
40  player_out.1          3521 non-null    object
41  fielders_name.1       4314 non-null    object
42  substitute            173 non-null     object
43  powerplays_from.1     126543 non-null  float64
44  to.1                  126543 non-null  float64
45  type.1                123685 non-null  object
46  target_overs          126543 non-null  float64
47  runs                  126543 non-null  float64
48  matchId               126855 non-null  int64
49  outcome_by_wickets    336 non-null     float64
50  byes                  477 non-null     float64
51  match_number          46094 non-null   float64
52  outcome_winner        48093 non-null   object
53  by_wickets            25672 non-null   float64
54  wickets_player_out    2397 non-null    object
55  kind                  2397 non-null    object
56  non_boundary          23 non-null      object
57  by_runs               178 non-null     float64
58  extras_legbyes        1350 non-null    float64
59  wides                 1980 non-null    float64
60  extras_legbyes.1      203 non-null     float64
61  wides.1               327 non-null     float64
62  byes.1                72 non-null      float64
63  extras_noballs        182 non-null     float64
64  replacements_role_in  45 non-null      object
65  role                  48 non-null      object
66  outcome_eliminator    8 non-null       object
67  result                8 non-null       object
68  extras_byes           77 non-null      float64
69  wickets_player_out.1  2257 non-null    object
70  kind.1                2257 non-null    object
71  stage                 62 non-null      object
72  method                21 non-null      object
73  review_by             426 non-null     object
74  umpire                432 non-null     object
75  batter                426 non-null     object
76  decision              426 non-null     object
77  umpires_call          85 non-null      object
78  review_by.1           154 non-null     object
79  umpire.1              154 non-null     object
80  batter.1              154 non-null     object
81  decision.1            154 non-null     object
```

```
match_data = ipl_data.iloc[:,11:49]
match_data.info()
```

```
#   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   innings_team         126855 non-null  object
 1   overs_over           20280 non-null   float64
 2   deliveries_batter    126001 non-null  object
 3   bowler               126001 non-null  object
 4   non_striker          126001 non-null  object
 5   runs_batter          126001 non-null  float64
 6   extras               126001 non-null  float64
 7   total                126001 non-null  float64
 8   extras_wides         3753 non-null    float64
 9   legbyes              1637 non-null    float64
 10  noballs              695 non-null     float64
 11  wickets_kind         3866 non-null    object
 12  player_out           3866 non-null    object
 13  fielders_name        4840 non-null    object
 14  powerplays_from      126855 non-null  float64
 15  to                   126855 non-null  float64
 16  type                 121766 non-null  object
 17  innings_2_team       123687 non-null  object
 18  overs_over.1         19128 non-null   float64
 19  deliveries_batter.1  117445 non-null  object
 20  bowler.1             117445 non-null  object
 21  non_striker.1        117445 non-null  object
 22  runs_batter.1        117445 non-null  float64
 23  extras.1             117445 non-null  float64
 24  total.1              117445 non-null  float64
 25  extras_wides.1       1585 non-null    float64
 26  noballs.1            137 non-null     float64
 27  legbyes.1            558 non-null     float64
 28  wickets_kind.1       3521 non-null    object
 29  player_out.1         3521 non-null    object
 30  fielders_name.1      4314 non-null    object
 31  substitute           173 non-null     object
 32  powerplays_from.1    126543 non-null  float64
 33  to.1                 126543 non-null  float64
 34  type.1               123685 non-null  object
 35  target_overs         126543 non-null  float64
 36  runs                 126543 non-null  float64
 37  matchId              126855 non-null  int64
```

```
# Summary statistics for numeric columns
numeric_summary = match_data.describe()
print(numeric_summary)
```

```
  overs_over    runs_batter         extras         total  \
count  20280.000000  126001.000000  126001.000000  126001.000000
mean       9.441568       1.262323       0.067317       1.329640
std        5.750996       1.629701       0.341158       1.616893
min        0.000000       0.000000       0.000000       0.000000
25%        4.000000       0.000000       0.000000       0.000000
50%        9.000000       1.000000       0.000000       1.000000
75%       14.000000       1.000000       0.000000       1.000000
max       19.000000       6.000000       5.000000       7.000000

       extras_wides      legbyes     noballs  powerplays_from             to
\
count   3753.000000  1637.000000  695.000000     1.268550e+05  126855.000000
mean       1.211031     1.270006    1.025899     1.000000e-01       5.604255
std        0.806897     0.804426    0.250457     2.255982e-13       0.200053
min        1.000000     1.000000    1.000000     1.000000e-01       1.600000
25%        1.000000     1.000000    1.000000     1.000000e-01       5.600000
50%        1.000000     1.000000    1.000000     1.000000e-01       5.600000
75%        1.000000     1.000000    1.000000     1.000000e-01       5.600000
max        5.000000     5.000000    5.000000     1.000000e-01       5.900000

       overs_over.1  ...   target_overs           runs       matchId  \
count  19128.000000  ...  126543.000000  126543.000000  1.268550e+05
mean       9.037171  ...      19.803194     164.602831  8.697117e+05
std        5.603415  ...       1.418912      31.724601  3.533232e+05
min        0.000000  ...       5.000000      43.000000  3.359820e+05
25%        4.000000  ...      20.000000     146.000000  5.483140e+05
50%        9.000000  ...      20.000000     165.000000  8.298170e+05
75%       14.000000  ...      20.000000     186.000000  1.216507e+06
max       19.000000  ...      20.000000     264.000000  1.370353e+06

             byes     byes.1  extras_legbyes  extras_legbyes.1        wides
\
count  477.000000  72.000000     1350.000000        203.000000  1980.000000
mean     1.781971   1.861111        1.352593          1.315271     1.198485
std      1.269656   1.314122        0.912871          0.843797     0.772505
min      1.000000   1.000000        1.000000          1.000000     1.000000
25%      1.000000   1.000000        1.000000          1.000000     1.000000
50%      1.000000   1.000000        1.000000          1.000000     1.000000
75%      2.000000   4.000000        1.000000          1.000000     1.000000
max      4.000000   4.000000        5.000000          4.000000     5.000000

          wides.1  extras_noballs
count  327.000000      182.000000
mean     1.165138        1.005495
std      0.643711        0.074125
min      1.000000        1.000000
25%      1.000000        1.000000
50%      1.000000        1.000000
75%      1.000000        1.000000
max      5.000000        2.000000
```

```
[8 rows x 28 columns]
```

Univariate Analysis

```
# Histograms for numerical columns
numeric_columns = match_data.select_dtypes(include='number').columns
for col in numeric_columns:
    plt.figure(figsize=(8, 6))
    plt.hist(ipl_data[col].dropna(), bins=20)
    plt.title(f'Histogram of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.show()

# Count plots for categorical columns
# categorical_columns = ['city', 'winner', 'venue', 'innings_team', ]  # List
all categorical columns
for col in categorical_columns:
    plt.figure(figsize=(10, 6))

    # Compute value counts for the column
    value_counts =
match_data[col].value_counts().sort_values(ascending=False)

    # Select the top 10 categories
    top_10 = value_counts.head(10)

    # Plot the top 10 categories
    sns.barplot(x=top_10.index, y=top_10.values)
    plt.title(f'Top 10 Categories in {col}')
    plt.xlabel(col)
    plt.ylabel('Count')
    plt.xticks(rotation=45)
    plt.show()
```

Bivariate Analysis

```
# Grouping data by 'deliveries_batter' and summing 'runs_batter'
runs_by_batter =
match_data.groupby('deliveries_batter')['runs_batter'].sum().sort_values(asce
nding=False)

# Plotting the aggregated runs for each batter
plt.figure(figsize=(12, 8))
runs_by_batter.head(10).plot(kind='bar', color='skyblue')
```

```
plt.xlabel('Batters')
plt.ylabel('Total Runs Scored')
plt.title('Total Runs Scored by Each Batter in 1st innings')
plt.xticks(rotation=45)
plt.show()
```

References

Websites

1. CRICSHEET. https://cricsheet.org/
2. ESPN. https://www.espncricinfo.com/