Predicting the final flourish : An IPL scorecard simulator using Machine Learning

Aabha Jahagirdar, Darshan Awale, Ganesh Pagadyakul, Mrunal Thamake

Computer Depatment,A.G. Patil Polytechnic Institute  
Opp SRP Camp, Vijapur Road, Solapur

[1AabhaJahagirdar@gmail.com](mailto:1AabhaJahagirdar@gmail.com)

2[darshanawale02@gmail.com](mailto:darshanawale02@gmail.com)

3[ganeshpagadyakul2616@gnail.com](mailto:ganeshpagadyakul2616@gnail.com)

[4mrunalthamake9@gmail.com](mailto:4mrunalthamake9@gmail.com)

Abstract— Fans all over the world are captivated by the intense cricket competition known as the Indian Premier League (IPL). For fanatics, there's an additional level of thrill when they can predict the result of these contests. In order to anticipate IPL match scores, this project explores the field of cricket analytics by creating a model based on Python.

The model makes use of numerous variables that have a big impact on how a match plays out. Among these are:

Team Dynamics: A significant factor in deciding the ultimate score is the past results of the bowling and batting teams.

Pitch Conditions: A pitch's ability to support batting or bowling can have a big impact on how many runs are scored.

Match Progression: The pressure on the batting team and scoring rate are reflected in the current over (beginning innings, middle overs, and death overs).

Wicket Loss Instinct.

The model seeks to understand the intricate relationships between different game elements and how they affect the overall score by examining five crucial elements. This research offers a data-driven method for predicting IPL match scores, adding to the expanding field of sports analytics.

Keywords—Machine Learning, sports analytics, linear regression, First Innings, IPL score prediction

1. Introduction

The excitement of picking the outcome of your most anticipated IPL match can only be imagined! This project uses machine learning to create a Python-based model for IPL score prediction, bringing that thrill one step closer.

Computers can learn through machine learning, and in this case, the data is all about previous Indian Premier League matches. Information on the teams involved, the kind of surface, and even the number of wickets taken so far will be fed into the algorithm.Page Layout.

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

We'll be employing the specific machine learning method known as linear regression. This facilitates the model's ability to discern connections between the final score and these other elements. The algorithm will learn to "predict" the score for a new match based on comparable circumstances by examining trends in previous matches.

This study explores the topic of cricket analytics and provides a data-driven approach to comprehend what really affects the game's ultimate result. So grab a seat, and let's examine the fascinating field of machine learning-based IPL match prediction.

1. Data Resources

Gathering pertinent information is the first stage in forecasting cricket matches. Numerous sources of data are available, such as player statistics, weather reports, historical match data, and live match feeds. Data from previous matches can offer insightful information on player statistics and team performance. A player's performance in prior games can be inferred from player statistics like batting and bowling averages. Weather-related information, such humidity and temperature, can also affect how a match turns out also the pitch conditions also affect the final result of the match. A key component of creating a successful ML model for cricket match prediction is choosing the right dataset. When choosing a relevant dataset for machine learning-based cricket match prediction, keep the following points in mind:

* Data Quality: The dataset should be of high quality, with minimal errors and missing values. The data should be obtained from reliable sources, such as official cricket websites, and should be thoroughly cleaned and pre-processed before being used for training the AI model.
* Data Availability: The dataset should contain a sufficient number of observations to train the AI model effectively. Ideally, the dataset should cover a wide range of matches, including different formats of the game (e.g., Test, ODI, T20), different teams, and different locations.
* Relevant Features: The dataset should contain relevant features that are known to impact the outcome of a cricket match. Some of the commonly used features for cricket match prediction include team performance indicators (e.g., win-loss record, batting and bowling averages), player statistics (e.g., batting and bowling averages, strike rates), weather conditions (e.g., temperature, humidity), pitch conditions (e.g., type of pitch, grass cover), and team composition (e.g., player rankings, injuries).
* Data Granularity: The dataset should have the appropriate granularity to support the prediction task. For example, if the prediction task is to predict the outcome of a match based on the first innings, then the dataset should contain features that reflect the first innings' performance, such as the total runs scored, wickets taken, and run rate.
* Data Diversity: The dataset should be diverse and representative of the different factors that impact the outcome of a cricket match. For example, the dataset should cover matches played in different weather conditions, different locations, and different levels of competition.

Overall, selecting a suitable dataset for cricket match prediction using ML requires careful consideration of several factors, including data quality, availability, relevance, granularity, and diversity. By selecting a suitable dataset, ML models can be trained effectively to predict the outcome of cricket matches accurately.

1. Preprocessing

The next step in predicting cricket matches is to pre-process the data. Data pre-processing involves cleaning the data, handling missing values, and transforming the data into a format suitable for machine learning algorithms. Cleaning the data involves removing irrelevant data and correcting errors. Handling missing values involves filling in missing values with appropriate values, such as the mean or median value. Transforming the data into a suitable format involves converting categorical data into numerical data and scaling the data to a common range. Here's an example of a cricket dataset that can be used to train AI models for predicting cricket match outcomes:

* Match details: This includes information about the match, such as the date, venue, and format (e.g., Test, ODI, T20).
* Team statistics: This includes information about the performance of the two teams, such as the win-loss record, batting and bowling averages, and run rate.
* Player statistics: This includes information about the performance of individual players, such as batting and bowling averages, strike rates, and player rankings.
* Pitch conditions: This includes information about the pitch conditions, such as the type of pitch, grass cover, and weather conditions.
* Toss details: This includes information about which team won the toss and chose to bat or bowl first.
* Inning details: This includes information about each team's innings, such as the total runs scored, wickets taken, and run rate.
* Outcome details: This includes information about the outcome of the match, such as which team won, the margin of victory, and the number of overs remaining.

1. Dataset Features

We are employing an ML-based strategy. Hence, a dataset, algorithmic training of that dataset, and model testing are the fundamental needs of a machine learning algorithm. Thus, we have imported the Kaggle dataset. Later on, using Linear Regression for scoring prediction, the accuracy was calculated and improved.

Score Prediction :- For this project, we have collected data of all the IPL matches played from 2008 to 2017. The dataset consists of 76015 numbers of rows. Dataset consists 15 columns over which we applied feature selection techniques and selected 8 features in which all 8 are input feature. The attributes selected were bat team, bowl team, type of pitch overs, runs, wickets, runs in previous 5 overs, wickets in previous 5 overs for score prediction.

TABLE   
Dataset Attributes and their values

|  |  |
| --- | --- |
| Attributes | Values |
| Batting team | Batting team names among 8 names in IPL |
| Bowling team | Bowling team names among 8 names in IPL |
| Pitch | The type of pitch among 6 types of pitches |
| Overs | Value > 5 over |
| Runs | 0-770 |
| Wickets | 0-10 |
| Runs scored in previous 5 overs | 0-770 |
| Wickets lost in previous 5 overs | 1-10 |
| Total runs | 0-770 |

All title and author details must be in single-column format and must be centered.

Every word in a title must be capitalized except for short minor words such as “a”, “an”, “and”, “as”, “at”, “by”, “for”, “from”, “if”, “in”, “into”, “on”, “or”, “of”, “the”, “to”, “with”.

Author details must not show any professional title (e.g. Managing Director), any academic title (e.g. Dr.) or any membership of any professional organization (e.g. Senior Member IEEE).

To avoid confusion, the family name must be written as the last part of each author name (e.g. John A.K. Smith).

Each affiliation must include, at the very least, the name of the company and the name of the country where the author is based (e.g. Causal Productions Pty Ltd, Australia).

Email address is compulsory for the corresponding author.

1. Section Headings

No more than 3 levels of headings should be used. All headings must be in 10pt font. Every word in a heading must be capitalized except for short minor words as listed in Section III-B.

1. Level-1 Heading: A level-1 heading must be in Small Caps, centered and numbered using uppercase Roman numerals. For example, see heading “III. Page Style” of this document. The two level-1 headings which must not be numbered are “Acknowledgment” and “References”.
2. Level-2 Heading: A level-2 heading must be in Italic, left-justified and numbered using an uppercase alphabetic letter followed by a period. For example, see heading “C. Section Headings” above.
3. Level-3 Heading: A level-3 heading must be indented, in Italic and numbered with an Arabic numeral followed by a right parenthesis. The level-3 heading must end with a colon. The body of the level-3 section immediately follows the level-3 heading in the same paragraph. For example, this paragraph begins with a level-3 heading.
4. Figures and Tables

Figures and tables must be centered in the column. Large figures and tables may span across both columns. Any table or figure that takes up more than 1 column width must be positioned either at the top or at the bottom of the page.

Graphics may be full color. All colors will be retained on the CDROM. Graphics must not use stipple fill patterns because they may not be reproduced properly. Please use only *SOLID FILL* colors which contrast well both on screen and on a black-and-white hardcopy, as shown in Fig. 1.



Fig. A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

Fig. 2 shows an example of a low-resolution image which would not be acceptable, whereas Fig. 3 shows an example of an image with adequate resolution. Check that the resolution is adequate to reveal the important detail in the figure.

Please check all figures in your paper both on screen and on a black-and-white hardcopy. When you check your paper on a black-and-white hardcopy, please ensure that:

* the colors used in each figure contrast well,
* the image used in each figure is clear,
* all text labels in each figure are legible.

1. Figure Captions

Figures must be numbered using Arabic numerals. Figure captions must be in 8 pt Regular font. Captions of a single line (e.g. Fig. 2) must be centered whereas multi-line captions must be justified (e.g. Fig. 1). Captions with figure numbers must be placed after their associated figures, as shown in Fig. 1.



Fig. Example of an unacceptable low-resolution image



Fig. Example of an image with acceptable resolution

1. Table Captions

Tables must be numbered using uppercase Roman numerals. Table captions must be centred and in 8 pt Regular font with Small Caps. Every word in a table caption must be capitalized except for short minor words as listed in Section III-B. Captions with table numbers must be placed before their associated tables, as shown in Table 1.

1. Page Numbers, Headers and Footers

Page numbers, headers and footers must not be used.

1. Links and Bookmarks

All hypertext links and section bookmarks will be removed from papers during the processing of papers for publication. If you need to refer to an Internet email address or URL in your paper, you must type out the address or URL fully in Regular font.

1. References

The heading of the References section must not be numbered. All reference items must be in 8 pt font. Please use Regular and Italic styles to distinguish different fields as shown in the References section. Number the reference items consecutively in square brackets (e.g. [1]).

When referring to a reference item, please simply use the reference number, as in [2]. Do not use “Ref. [3]” or “Reference [3]” except at the beginning of a sentence, e.g. “Reference [3] shows …”. Multiple references are each numbered with separate brackets (e.g. [2], [3], [4]–[6]).

Examples of reference items of different categories shown in the References section include:

* example of a book in [1]
* example of a book in a series in [2]
* example of a journal article in [3]
* example of a conference paper in [4]
* example of a patent in [5]
* example of a website in [6]
* example of a web page in [7]
* example of a databook as a manual in [8]
* example of a datasheet in [9]
* example of a master’s thesis in [10]
* example of a technical report in [11]
* example of a standard in [12]

1. Conclusions

The version of this template is V2. Most of the formatting instructions in this document have been compiled by Causal Productions from the IEEE LaTeX style files. Causal Productions offers both A4 templates and US Letter templates for LaTeX and Microsoft Word. The LaTeX templates depend on the official IEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templates are self-contained. Causal Productions has used its best efforts to ensure that the templates have the same appearance.

Acknowledgment

The heading of the Acknowledgment section and the References section must not be numbered.

Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

References

1. S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
2. J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
3. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.
4. M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in *Proc. ECOC’00*, 2000, paper 11.3.4, p. 109.
5. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.
6. (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
7. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
8. *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
9. “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.
10. A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
11. J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
12. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.