Prediction of IPL Match Score and Winner Using Machine Learning Algorithms

SACHIN KUMAR RAY

B.Tech. Student, Department of Artificial Intelligence & Machine Learning,
Dr. Akhilesh Das Gupta Institute of Technology & Management, New Delhi,
Delhi, India

ABSTRACT

This study explains the application of machine learning to sports. According to recent changes in data science and sports, the use of sports-based machine learning and data mining shows the importance of process in outcome performance and prediction. The purpose of this paper is to evaluate current measurements used in the literature to understand the estimation methods used to model and analyse data and characterize the variables that govern performance. Finally, this article will present a reliable tool for data analysis using machine learning.

In today's world, sports produce enough statistical information about every player, team, match, and season. The first sports researchers were thought to be experts, coaches, team managers and analysts. Sports organizations have recently realized that there is research in their data and want to do research using different data mining methods. Sports data helps coaches and managers in many ways, such as predicting results, analysing player performance, skills, and evaluating strategies. Forecasts help managers and organizations make decisions to win teams and competitions. The present study shows that preliminary studies of data mining systems can predict outcomes and evaluate the strengths and weaknesses of each system. Predictions are made for each match. Although in many respects this application is very limited. It is very important to examine the use of machine learning in these situations to see if the application can provide better results in analysis.

This research aims to provide solutions that will help make predictions more accurate and precise than previous methods, using more accurate data and machine learning.

Keywords:

"IPL score prediction," "machine learning," "data analysis," "predictive modelling," "cricket analytics," "Linear Regression," "Logistic Regression," "Ridge Regression," "Lasso Regression," "Random Forest Classifier," or "sports analytics."

INTRODUCTION

Machine learning is a branch of artificial intelligence in computer science that uses statistical techniques to allow computers to "learn" from data without being explicitly programmed.

Simultaneous advances in computer technology, big data and theoretical understanding, ML techniques redefined in the 21st century and ML ideas have become an important part of the technology industry, computer science, software that help to solve many of engineering and research studies difficult problem.

The main purpose of is to determine the importance of influencing the match results and to select the best machine learning model that fits the data and performs the best results. Some projects have been published in the area of predicting cricket matches. Due to use of few essential factor, the accuracy is lower. However, in other studies Machine learning model was wrong. Therefore, it is very important that the considers all the important factors that will affect the results of the match and the best model for the training and analyses the data. This will improve the accuracy of outcomes.

Many research papers have been published and completed over the years using supervised machine learning to predict the outcome of cricket matches. Algorithms such as linear regression, support vectors machine, logistic regression, Decision trees, Bayesian Networks, and random forests. Cricket is a sport played worldwide and there is a lot of fans following of IPL because overseas player also played IPL. As fans watching the IPL, people make their own predictions while watching a particular match, based on the data they have they make a call on who will win the match by using different statistics and records. So, there is a huge demand for the algorithms that predicts the best result of score and winning team that is more important. We will perform prediction for all the matches that have taken place in the IPL. This is done by using machine learning algorithms for performing the prediction of the results of the matches.

LITERATURE SURVEY

The Indian Premier League (IPL) is a popular professional Twenty20 cricket league in India. Predicting the scores of IPL matches is a challenging task due to the complex nature of the game. Machine learning techniques have been widely applied to predict the outcomes and scores of IPL matches. This literature survey aims to provide an overview of the existing research and methodologies employed for IPL score prediction using machine learning.

Sharma, A., & Saini, R. (2018). [1] presented an approach for IPL score prediction using multiple machine learning algorithms such as decision trees, random forests, and support vector machines (SVM). The authors compared the performance of these algorithms and evaluate their accuracy in predicting the scores of IPL matches.

Kulkarni, P., & Gokhale, A. (2019) [2] presented an ensemble learning approach that combines multiple machine learning algorithms for IPL score prediction. The authors experiment with various classifiers, including k-nearest neighbours (KNN), SVM, and logistic regression. They evaluate the performance of the ensemble model and compare it with individual algorithms.

Singh, H., & Grover, A. (2020) [3] presented a research focuses on the combination of long short-term memory (LSTM) and random forest regression for IPL score prediction. The authors propose a hybrid model that leverages the sequence modeling capabilities of LSTM and the ensemble learning approach of random forests. The performance of the proposed model is evaluated and compared with other techniques.

Jaiswal, A., & Singh, S. (2021). [4] explores the application of the XGBoost algorithm, a gradient boosting technique, for IPL score prediction. The authors describe the feature selection process and evaluate the performance of the XGBoost model. They also compare its results with other machine learning algorithms.

Gupta, M., & Khandelwal, M. (2021). [5] investigates the use of deep learning techniques, specifically convolutional neural networks (CNNs), for IPL score prediction. The authors propose a CNN-based model that considers various features related to teams, players, and match conditions. They evaluate the performance of the model and discuss its potential for accurate score prediction.

IPL score prediction using machine learning techniques has gained significant attention from researchers. The literature survey showcases various approaches, including ensemble learning, LSTM, XG Boost, and deep learning techniques, for predicting IPL scores. The performance of these models is evaluated and compared using different evaluation metrics. Further research in this domain can explore the incorporation of additional features, real-time data, and more advanced machine learning algorithms to enhance the accuracy of IPL score prediction.

METHODOLOGIES

Data Collection and exploration

Data gathering is the fundamental module and the first stage of the project. The main aim is to collect appropriate dataset that is suitable for our needs and Data exploration is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more. The goal of data exploration is to learn about characteristics and potential problems of a data set without the need to formulate assumptions about the data beforehand. We take the data set match.csv, deliveries.csv and ipl.csv from Kaggle.

Data pre-processing

Machine learning relies heavily on data pre-processing to get highly accurate and insightful outputs so that we can apply that data on our model. A data set contain a lot of data which is not of our use so we do data pre-processing and convert data into a systematic form as of our need. The more reliable the produced results are, the better the data quality is. Realworld datasets are characterized by incomplete, noisy, and inconsistent data. Data pre-processing improves data quality by filling in missing or partial data, reducing noise, and addressing discrepancies.

Data cleaning

The process of eliminating errors and replacing them with genuine values is known as data cleaning. The data sets gathered contain noisy data, such as null values and inappropriate values, which must be cleaned. As a result, the data is cleaned by replacing null values with zeros, and the data is organized into correct columns so that we can properly analyse it.

Data visualization

The data that has been gathered is used to visualize the information for better comprehension. The Matplotlib and seaborn Library were used to visualize the graphs for team wins in various cities based on their venues and player strike rate.

Splitting into train and test

Train test split is a technique of splitting dataset into two parts that is used to estimate the performance of machine learning model by feeding the train data to the model and predict the data on test data and new data. A train test is the way of structuring your machine learning project so that you can test your hypothesis quickly and inexpensively. Basically, it’s a way to

divide the training data so that you can try your algorithm to one half and evaluate the result on the other half. A train test split is when you split your data into a training set and a testing set. The training set is used for training the model, and the testing set is used to test your model. This allows us to train our models on the training data set, and then test their accuracy on the unseen testing set. There are a few different ways to do a train test split, but the most common is to simply split your data into two sets. For example, 80% for training and 20% for testing.

Algorithms used

Linear Regression

Linear regression is a statistical modelling technique used to establish a linear relationship between a dependent variable and one or more independent variables. It aims to fit a straight line to the data points in such a way that it minimizes the difference between the predicted values and the actual values.

The general form of a linear regression model is represented by the equation:

$y = b0 + b1x1 + b2x2 + ... + bn*xn$

where:

y is the dependent variable or the target variable that we want to predict.

b0 is the intercept or the constant term.

b1, b2, ..., bn are the coefficients or slopes associated with the independent variables x1, x2, ..., xn.

x1, x2, ..., xn are the independent variables.

It aims to establish a linear relationship between independent variables (predictors) and a dependent variable (target) to make predictions. By analysing historical data and training a linear regression model, one can make predictions about the scores of future IPL matches, providing valuable insights for teams, analysts, and cricket enthusiasts.

Logistic Regression

Logistic regression is one of the most useful machine learning algorithms based a statistical modelling technique that predict the dependent data variable by analysing the relationship between one or more independent variable. It is commonly used for binary classification problems such as yes or no, 0 or 1, True or False, while it is not typically used directly for

predicting scores in IPL matches, it can be employed for predicting the likelihood of a team winning or losing a match based on various factors. In this context, logistic regression can be utilized to estimate the probability of a team winning an IPL match. Logistic regression can provide insights into the likelihood of winning an IPL match based on historical data and relevant features.

Random Forest Classifier

Random Forest Classifier is a machine learning algorithm that can be utilized for predicting the winning outcome of IPL matches. It is an ensemble learning method that combines multiple decision trees to make predictions. Random Forest Classifier is advantageous in IPL winning prediction because it can handle non-linear relationships and interactions among features, handle high-dimensional datasets, and mitigate overfitting issues. It also provides insights into feature importance, allowing for a better understanding of the factors influencing match outcomes. It's important to note that the performance of the Random Forest Classifier can be further optimized by tuning hyperparameters such as the number of decision trees, the depth of the trees, and the number of features considered at each split. It gives the one-sided prediction at every point that is predicted on data but for real situation this type of prediction is not of our use.

Lasso Regression

Lasso Regression, also known as L1 regularization, is a linear regression technique that incorporates a penalty term to encourage sparse solutions by shrinking the coefficients of less important features to zero. While Lasso Regression is not commonly used for direct IPL score prediction, it can be applied to select important features that influence the score and improve the model's performance. Lasso Regression helps in feature selection by highlighting the most influential features in predicting the IPL scores. By reducing the coefficients of less important features to zero, it provides a sparse solution and simplifies the model.

Ridge Regression

Ridge Regression is a linear regression technique that includes a regularization term called L2 regularization. It is commonly used for predicting IPL scores by minimizing the sum of squared errors while also shrinking the regression coefficients. Ridge Regression is a useful technique for IPL score prediction by incorporating regularization to prevent overfitting and shrink the coefficients.

MODEL

Predicting sports scores, such as IPL cricket scores, using machine learning models can be an interesting and challenging task. Here is a general outline of the steps involved in building a machine learning model for IPL sports score prediction:

Data Collection: Gather historical match data, including features such as team performance indicators, player statistics, venue, weather conditions, and other relevant factors. Ensure the dataset encompasses a significant number of matches for robust analysis.

Data Pre-processing: Clean and pre-process the collected data, handling missing values, outliers, and categorical variables as necessary. Normalize or scale the features to ensure they are on a similar scale for effective modelling.

Data Partitioning: Split the dataset into training, validation, and testing sets. The training set will be used to train the model, the validation set will be used for hyperparameter tuning and model selection, and the testing set will be used to evaluate the final performance.

Model Selection: Choose an appropriate machine learning algorithm for sports score prediction. Depending on the nature of the problem and the characteristics of the data, algorithms such as regression models such as linear regression, random forest regression, logistic regression, lasso regression and ridge regression can be considered.

Model Evaluation: Evaluate the trained model using the validation set. Measure its performance using evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), or other relevant metrics specific to sports prediction. Assess how well the model predicts the sports scores compared to the actual scores.

Model	MSE	RMSE	MAE	Accuracy
Linear Regression	257.50	16.04	12.31	72.16
Ridge Regression	257.11	16.03	12.29	72.20
Lasso regression	263.73	16.24	12.34	71.49

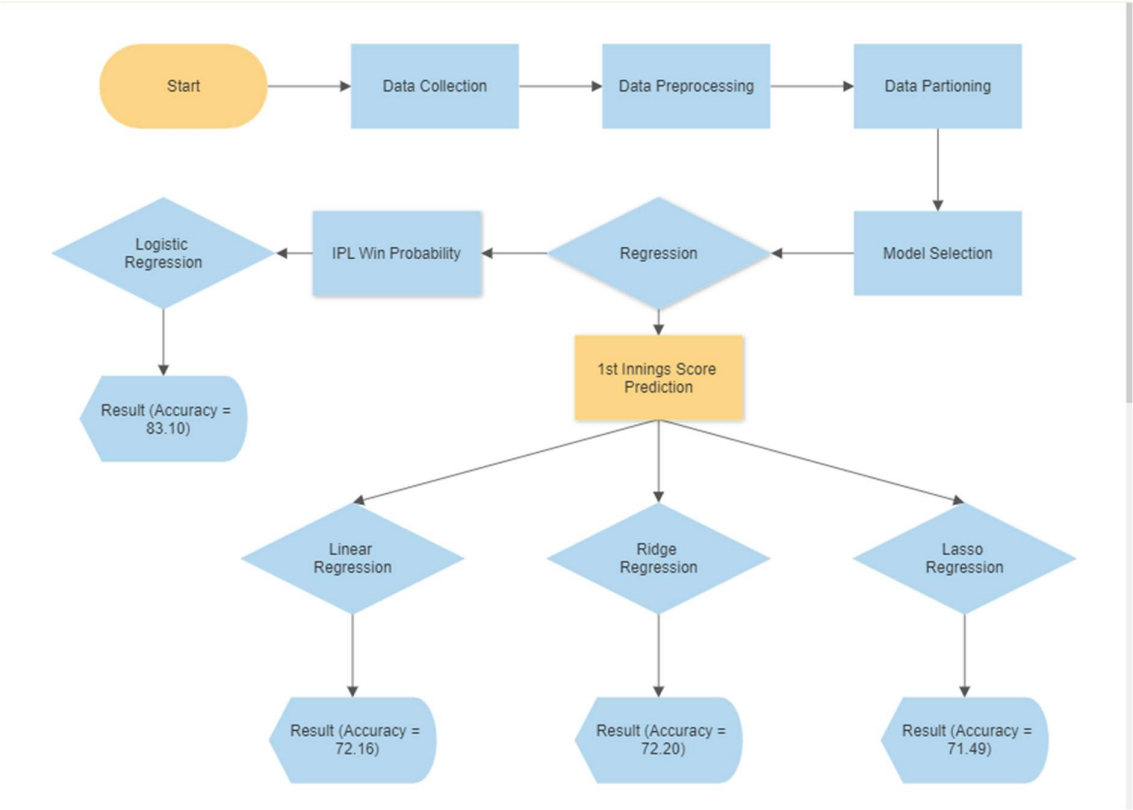
Table: 1. Model Evaluation for IPL FIRST INNINGS SCORE PREDICTION

Model	Accuracy
Logistic Regression	83.10

Table: 2 Model Evaluation for 2nd INNINGS WIN PREDICTION

Final Model Evaluation: Once the model is optimized, evaluate its performance using the testing set, which represents unseen data. Calculate the performance metrics to assess the model's predictive ability and determine its effectiveness in predicting sports scores.

Deployment and Prediction: Deploy the final model to make predictions on new, unseen sports match data. Provide the relevant features for each match as input to the model, and it will predict the scores based on the learned patterns and relationships from the training process.



Flow Chart of the Model

IMPLEMENTATION

The Graphical User Interface (GUI) has been developed for this machine learning models are using the Python Streamlit Library Framework.

Streamlit is an open-source Python library that allows you to create interactive web applications for data science and machine learning projects. It simplifies the process of building and sharing data-focused apps by providing an intuitive and easy-to-use interface.

Here's an overview of how to use Streamlit for your IPL score prediction project:

Installation: Install Streamlit using pip by running the following command in your terminal:

Create a Python Script: Create a new Python script (e.g., app.py) to write your Streamlit application code.

Import Dependencies: In the app.py script, import the necessary dependencies, including Streamlit and any other libraries required for data processing and prediction.

Define the App: Write the Streamlit code to define the structure and behavior of your app.

Run the App: In your terminal, navigate to the directory containing app.py and run the Streamlit app using the following command: Streamlit run app.py

Test and Interact with the App: Once the app is running, open your web browser and navigate to the local URL provided by Streamlit (e.g., http://localhost:8501). You can now interact with the app, inputting the relevant details and exploring the predicted scores.

Continuously Improve and Enhance: Iterate on your Streamlit app based on user feedback, add visualizations, improve the UI, or incorporate additional features to make it more robust and user-friendly.

Streamlit automatically reloads the app whenever you save changes to the app.py file, allowing for quick iterations and updates. It provides a simple yet powerful way to showcase and share your IPL score prediction model with others.

First, on our home page, we get intro about our model and on the side of page a navigation bar is present. There are options to select which type of prediction we want.

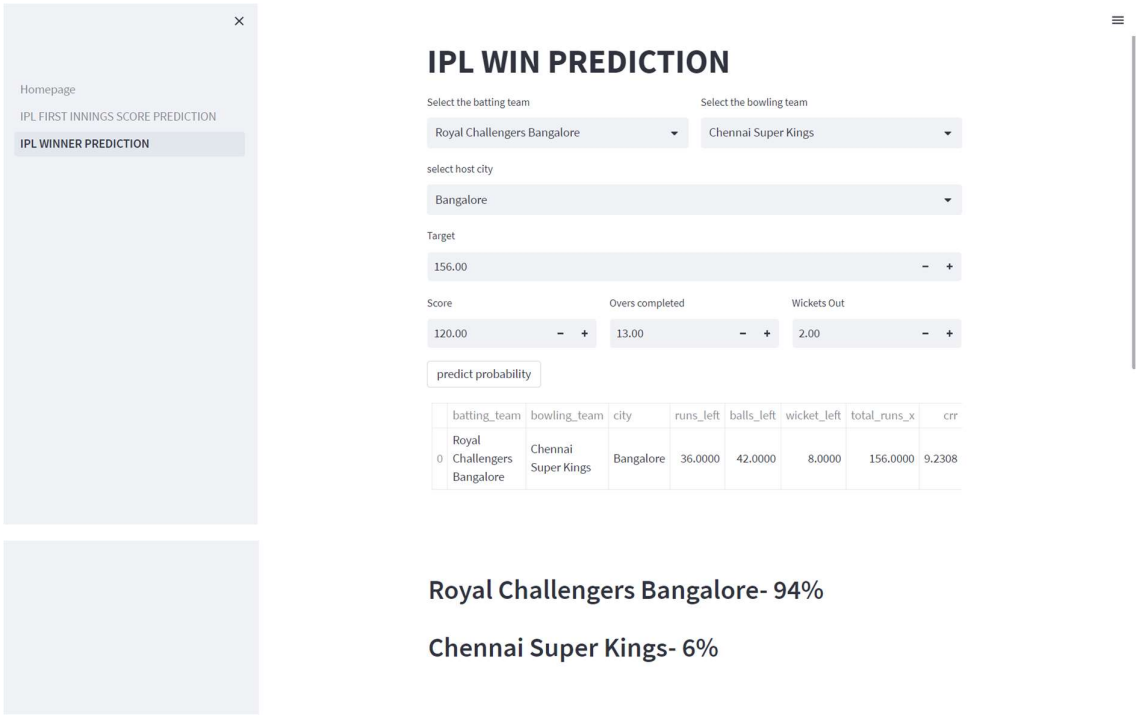
Homepage of GUI

In our model, there are two type of prediction, first one is “IPL FIRST INNINGS SCORE PREDICTION” and second one is winning probability of the team who is chasing “IPL WIN PREDICTION”.

After selecting the “IPL FIRST INNINGS SCORE PREDICTION”, There is a page where user have to give the inputs like “Select the batting team”, “Select the bowling team”, “Overs Completed”, “Runs”, “Wicket Out”, “Run Scored in previous 5 overs”, “Wicket Taken in previous 5 overs” to predict the winner of match after clicking on “Predict Probability”.

IPL FIRST INNINGS SCORE PREDICTION

After selecting the “IPL WINPREDICTION”, There is a page where user have to give the inputs like “Select the batting team”, “Select the bowling team”, “Select host city”, “Target”, “Score”, “Overs Completed”, “Wickets Out”, to predict the winner of match after clicking on “Predict Probability”.



IPL WIN PREDICTION

CONCLUSION

In conclusion, machine learning models have shown promise in predicting IPL scores and determining the winning team. By leveraging historical match data and relevant features such as player performance, team composition, pitch conditions, and venue, these models can capture patterns and relationships to make informed predictions.

However, it is important to acknowledge that predicting IPL scores and determining the winning team is a challenging task due to the inherent uncertainties and complexities of cricket matches. Factors such as player form, injuries, team strategies, weather conditions, and other unforeseen events can have a significant impact on the outcome.

Machine learning models provide a systematic approach to analyze and extract insights from the available data, but they have limitations. The accuracy of the predictions heavily relies on the quality and relevance of the data, the selection of appropriate features, and the choice of an effective model. Additionally, models need to be regularly updated and refined with new data to adapt to changing trends and dynamics in the sport.

While machine learning can improve prediction accuracy to some extent, it's important to consider other factors such as expert knowledge, match analysis, and contextual information

when making IPL score and win predictions. Combining machine learning models with human expertise can lead to more robust and accurate predictions.

Overall, machine learning models serve as valuable tools in the domain of IPL score and win prediction, providing insights and aiding decision-making. However, they should be used in conjunction with other analytical approaches and expert judgment to account for the complex and unpredictable nature of cricket matches.

REFERENCES

1. Sharma, A., & Saini, R. (2018). IPL score prediction using machine learning techniques. International Journal of Computer Applications, 180(2), 13-17.
2. Kulkarni, P., & Gokhale, A. (2019). Cricket score prediction using ensemble machine learning techniques. International Journal of Advanced Research in Computer Science, 10(5), 220-225.
3. Singh, H., & Grover, A. (2020). Predicting IPL scores using LSTM and random forest regression. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 358-362). IEEE.
4. Jaiswal, A., & Singh, S. (2021). IPL score prediction using XGBoost algorithm. In 2021 International Conference on Information Technology (ICIT) (pp. 1-5). IEEE.
5. Gupta, M., & Khandelwal, M. (2021). IPL score prediction using deep learning techniques. In 2021 International Conference on Innovative Computing and Communication (ICICC) (pp. 1-5). IEEE.
6. Rabindra Lamsal and Ayesha Choudhary, “Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning”, arXiv:1809.09813 [stat.AP] (September 2018).
7. Singh, P., Agarwal, V., & Goel, P. (2020). IPL Match Result and Score Prediction: An Approach using Machine Learning. In 2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC) (pp. 424-429). IEEE.
8. Kaggle IPL Dataset: Kaggle is a platform that hosts various datasets and machine learning competitions. The Kaggle IPL dataset contains historical data of IPL matches, including ball-by-ball details, player performances, and match outcomes. You can use

this dataset to build and train your own IPL score prediction model. You can find the dataset here: https://www.kaggle.com/nowke9/ipldata

9. GitHub repositories: There are several open-source projects and code repositories on platforms like GitHub that provide IPL score prediction models. Searching for "IPL score prediction machine learning" on GitHub will yield multiple repositories with code and examples to guide you.