

Machine Learning in Match Prediction

Winner Prediction of ipl

**BY:**

Abhishek Mane (20bds001)

Amit Kumar Verma (20bds005)

Ashutosh Rana (20bds011)

Gagan Vadlamudi (20bds019)

Suyash Nishikant Kamble (20bds029)

Praveen Gundyagol (20bds042)

Surasani Venkata Amarnadh Reddy (20bds055)

Karuna Prakash Veepuri (20bds062)

Yogesh Parashram Shinde (20bds065)

Progressors | ML in Match Prediction | May 9, 2021

# What is Machine learning?

Machine learning is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence.

In simple Words,

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

Use of Machine Learning in different sections of Sports:

Machine learning in sport is continuously used for the most varied uses, creating a branch in its own right like Sports analytics. Sports Analytics differs in various areas such as

Sports Analytics is the process that identifies and acquires knowledge and insight into the performance of potential players based on the use of a variety of data sources such as game data and individual player performance data. This advanced and sophisticated type of analysis should be able to extract valuable usable information that coaches can use.

Sports Analytics can be used in infinite ways including:

* **Predict the outcome of a game**
* **Predict the performance of teams or individual players**
* **Creation of new strategies for upcoming competitions**



Machine Learning in IPL match Prediction:

In Machine Learning, the problems are categorized into 2 groups mainly: Regression Problem and Classification problem. The Regression problem deals with the kind of problems having continuous values as output while in the Classification problem the outputs are categorical values. Since the output of winner prediction is a categorical value, the problem which we are trying to solve is a Classification problem.

So where to start, and what to do?

* Understand the dataset.
* Clean the data.
* Analyze the candidate columns to be Features.
* Process the features as required by the model/algorithm.
* Train the model/algorithm on training data.
* Test the model/algorithm on testing data.
* Tune the model/algorithm for higher accuracy.

**Let’s get started!**

Abstract:

In cricket, particularly the twenty20 format is most watched and loved by the people, where no one can guess who will win the match until the last ball of the last over. In India, The Indian Premier League (IPL) started in 2008 and now it is the most popular T20 league in the world. So we decided to develop a machine learning model for predicting the outcome of its matches. Winning in a Cricket Match depends on many key factors like a home ground advantage, past performances on that ground, records at the same venue, the overall experience of the players, record with a particular opposition, and the overall current form of the team and also the individual player. This paper briefs about the key factors that affect the result of the cricket match and the regression model that best fits this data and gives the best predictions. Cricket, the mainstream and widely played sport across India which has the most noteworthy fan base. Indian Premier League follows 20-20 format which is very unpredictable. IPL match predictor is a ML based prediction approach where the data sets and previous stats are trained in all dimensions covering all important factors such as: Toss, Home Ground, Captains, Favorite Players, Opposition Battle, Previous Stats etc., with each factor having different strength.

Predicting IPL Winner:

Classification and Regression are the two branches of Supervised Learning in the field of Machine Learning. These are the basic topics that one should learn when starting their journey with Machine Learning. Doing projects is the only way through which one can learn and master these topics.

**Dataset:**

We have scraped the data from ICC’s T20 top 100 players and Cricbuzz websites using Beautiful Soup module and obtained the data for Batsmen, Bowler and All-Rounder separately.

**Features used for Batsmen:**

• Innings

• Runs Scored

• Batting Average

• Batting Strike Rate

• Fifties

• Fours

• Sixes

**Features used for Bowlers:**

• Innings

• Wickets

• Economy

• Bowling Average

• Bowling Strike Rate

**Features used for All-Rounders:**

• Innings as Batsmen

• Runs Scored

• Batting Average

• Batting Strike Rate

• Fifties •

Fours

• Sixes

• Innings as Bowler

• Wickets

• Economy

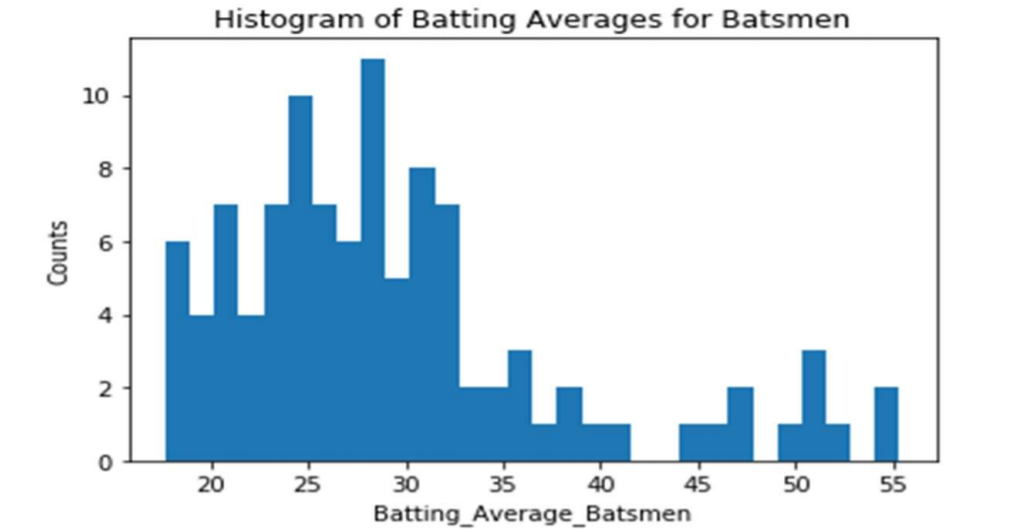
• Bowling Average

• Bowling Strike Rate

Now that we have our data, it is time for us to observe and understand the data.

**Data Pre-Processing:**

We have plotted the histogram of the data and observed that the data is highly right skewed. To overcome the skewness, we have log-transformed the data, so that the data is now approximately normal and our models work well.



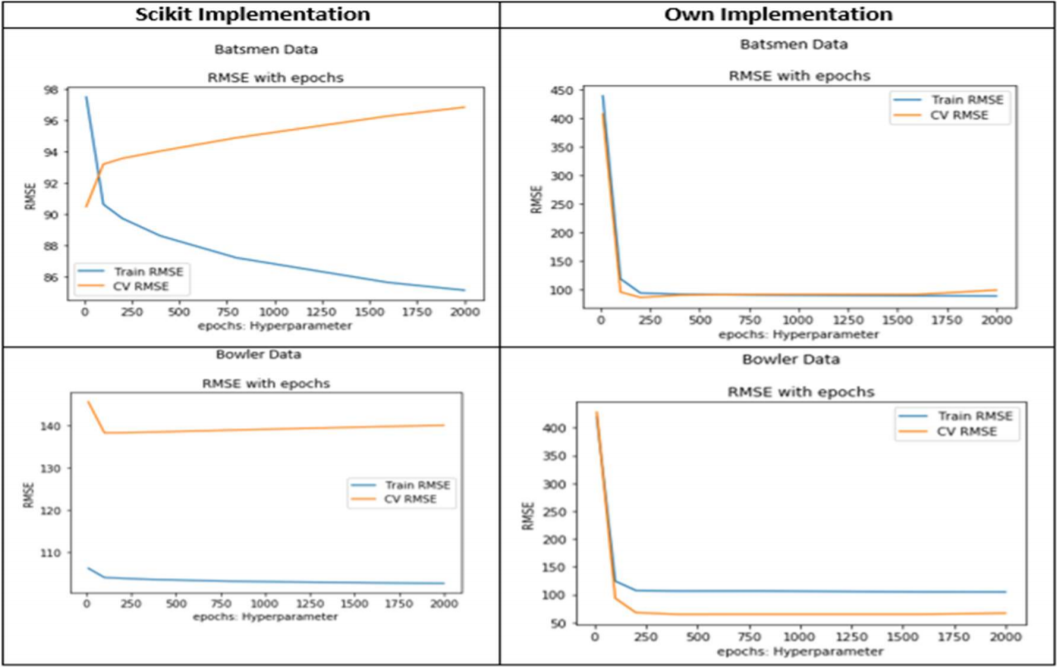
We have observed that some records have zero values for some of the features. We have dropped those records from our datasets. Most probably they are the players who played very less number of matches.

**Model Building:**

Now that we have processed and cleaned our data, we have to build the machine learning models on this cleaned data. We have implemented SGDRegressor, KNN-Regressor,Linear Regression using LeastSquare Estimates, Weighted KNN-Regressor and compared our models with the Scikit learn’s model and we have achieved almost similar results to that of the Scikit Implementation.

**Regression:**

We have provided the training data as the input to the models and found the best number of epoch using Cross Validation Data. AS we can see in the plots, the training error decreases as the epochs increases which means we are going towards convergence. We can increase the number of epochs to reduce the error furthermore, but we have stopped at 2000 epochs as the error fall rate is almost steady at this point. We have got the best number of epochs using cross validation. Now, we can run the model with this best number of epochs on test data and we can obtain the predicted values. We can then compare our predicted values with the actual values to get the error.



**Data Visualization:**

The following is the data depicting the various statistics related to the project which impacts the outcome of the winning team.

Conclusion:

From the study there are numerous elements which impact result of any IPL match is observed. Main factors that fundamentally impact any IPL match could be their host group, non-home group, arena, winner of toss and many more. This relatively helped in the calculation of strength. Different ML techniques were handed down for IPL data set which contributed to this study. The data set consists of all the IPL matches that were held from the past 6 years that is from 2014 to 2019. The prepared models were utilized to foresee the result of IPL matches. The T20 cricket has a scope for changeability, because even few balls can totally change the game. IPL was started 12 years back, there were very less number of games played compared to 50-50 and test games. Thus, structuring ML for anticipating game result with a precession of 75% is exceptionally good at this stage.