Predicting Scores, Wickets and Results of Cricket Matches using Machine Learning

Faizan Ahmad
faizann288@gmail.com
Crickytics

Abstract
Cricket is the second most loved and watched sports in the world. A lot of work is being done in sports analytics but cricket is a sport where there is little or no work being done in its analysis. This paper aims to apply machine learning methods to make various predictions in cricket that are result of the match and individual scores of players. We show that the use of machine learning in making cricket predictions outweighs the random guesses that people make.

Keywords - Cricket, Machine Learning, Prediction

1. INTRODUCTION

Cricket is the second most watched and loved sports in the world. With the advances in machine learning and data analytics, a lot of work is being done in sports like football, baseball etc but very little work is being done in applying machine learning to cricket. This paper aims to fill this gap. We apply machine learning methods to make two important predictions for a cricket match i.e the result of the match before it even starts and the individual scores of batsman.

We show that machine learning methods outperform random guessing. The accuracy of our model also outperforms the models that have been made before. We believe that this work is first of its kind and further work should be done in this field.

The paper is divided into sections. The first section describes some related work done in cricket using machine learning. The second section describes our problem statements. The third section describes our methodology of data collection, feature extraction and machine learning methods that we used. The fourth section describes our experimental set up and the last section concludes with a few recommendations for future work.

2. RELATED WORK

In the field of cricket, very little work has been done by applying machine learning methods to the field of cricket. Work done by [1],[2],[3] is the only work that has been done for cricket.

[1] has applied some analytics techniques to weigh the performance of bowlers. [2] has done extensive work in applying machine learning to English country cricket matches. [3] has done some work in predicting ODI matches.

3. PROBLEM STATEMENT

We have two problems of interest that we want to predict.

1. We want to predict the result of a match before it starts. This is a hard problem since the game of cricket is very random. This problem is a binary classification problem where 1 means team will win the match and 0 means team will lose the match.
2. We want to predict the individual scores of each player playing in the match. We made this problem as a 3 class classification problem where 3 classes are the score ranges that a player will make i.e the first class is score from 0 to 30, second is from 30 to 80 and third is 80+.
3. Our third problem statement is the most interesting. Given last 18 balls in a T20 match, we want to predict if there will be a wicket in the next 6 balls..

4. METHODOLOGY

This section describes in detail the methodology adopted to solve these above problems. In summary, we first scraped data from online websites including Cricinfo and Icc-cricket. We made json files from this scraped data and then converted these json files into a database. After our database was complete, we had two three tables; one for matches details, one for ball by ball details in each match and one for player names and ids. After our data collection step, we moved on to feature extraction for our model. We tried different sets of features and finalized around 15 features in total to use in all of our problems. After feature selection, we opted 10 different machine learning models and applied them all in our initial testing. After our initial testing, we removed some models and kept only a few that were giving us the best results. After all the pruning, we were left with one model for each problem. In the end, we computed accuracies and precisions.

4.1. Data Collection

. The first and longest step in our whole process was data collection. Data was collected from cricinfo and icc cricket's website. Both these websites provide json files for each match. These json files contain every single bit of information about a match. There was a lot of clutter and a lot of data processing had to be done. After the initial pre-processing, data was converted into another set of jsons which was then converted into a database. One thing to note here is that we have chosen one team as our source team and that is Pakistan. We only keep data for Pakistan and this is the team for which we are applying our machine learning models.

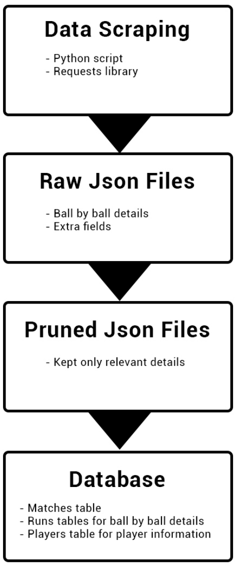

Figure 1. Data Collection

4.2. Feature Selection

The next step after data collection was feature selection. This is the most important part in any machine learning project. We tried various features but the accuracy was not affected by many features. In the end, we were only left with a few features that best described our data set. The features we used were.

1. **Predicting Match Outcome**
    - Mean of average of each batsman against the opposition for all the matches played before this match
    - Mean of average of each batsman against the opposition in the same country where match is being played for all the matches played before this match
    - Mean of wickets of each player against the opposition for all the matches played before this match
    - Mean of wickets of each player against the opposition in the same country where match is being played for all the matches played before this match
    - Win ratio of Pakistan against the opposition
    - Win ratio of Pakistan against the opposition at the same country where match is being played
2. **Predicting Player Scores**
    - Opposition Team (We give an integer value to each opposition team)
    - Whether the match is day and night or just day
    - Country in which match is being played
    - Ground (We assign an integer value to each ground)
3. **Predicting Next Wicket**
    - Score of each ball from last 18 balls.
    - Wicket (binary variable) for each ball in last 18 balls.
    - Overs going on during 18 balls.
    - Ball Speed of 18 balls.
    - Ball movement variation of 18 balls.
    - Playing Team
    - Opponent Team
    - Wickets till now
    - Country where match is being played

4.3. Algorithms

After we have our features, we apply machine learning algorithms. We used one algorithm for each task. We tried more than 10 different algorithms but ended up with only 2 of them since the others were not giving as good results as the ones we selected. Our algorithms for our problems were.

1. **Predicting Match Outcome.** We chose Naive Bayes for this classification problem. This problem deals a lot with independent probabilities. Therefore, naive bayes gives the best result.

2. **Predicting Player Scores.** We used Random Forests for this classification problem. Random forests give the best result on this problem along with decision trees. One thing to note here is that we get separate accuracies for each player.

3. **Predicting Next Wicket.** Since this problem involved a lot of features, we opted neural networks for this prediction problem. We used neural networks with 180 hidden layers divided into set of 60 for each layer. This was a binary classification problem where 0 means there will be no wicket in the next balls and 1 means there will be a wicket in the next six balls. One important thing about this problem is that we used **Soft Classification** in this problem. Soft classification means that we dealt with probabilities instead of raw accuracy from the model. If the probability of 0 was greater than 88 percent, we assigned it 0 otherwise we assigned 1. This was done after an extensive testing of probability values. This gave us much better accuracies and precisions.

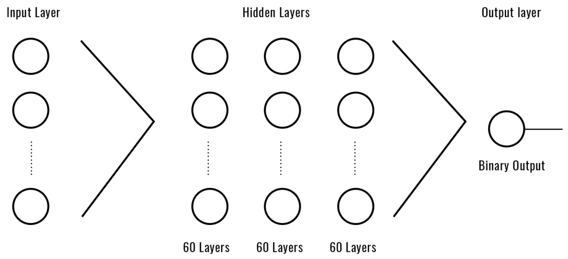

Figure 2. Neural Networks Architecture for wicket prediction

5. EXPERIMENTAL SETUP

After we are done with everything, we write code to see what accuracy do we get.

5.1. Predicting Match Outcome

For this, we got 184 ODI matches of Pakistan for the last 10 years from our database. We extracted features by writing a python script. Afterwards, we use machine learning models from scikit learn library. We applied various models and noted their results.

5.2. Predicting Player Scores

For this, we got 184 ODI matches of Pakistan for the last 10 years and score of each player in each match from our database. We extracted features by writing a python script. Afterwards, we use machine learning models from scikit learn library. We applied various models and noted their results.

5.3. Predicting Next Wicket

For this, we got 400 T20 International matches of all teams for the last 10 years and score of each player in each match from our database. We extracted features from the data where each feature was a vector containing all the information of our features and label was whether a wicket has actually fallen in the next six balls or not.

6. RESULTS

6.1. Predicting Match Outcome

The results of different algorithms were not very similar and different algorithms gave different accuracies. We selected the ones that gave the best accuracy. We applied the same algorithms using different training and testing sets and the algorithms that were giving high accuracy continued to give high accuracy. Therefore, we were sure that these are the algorithms that are best learning the patterns in the data.

Classification Algorithm	Accuracy	Precision
Naive Bayes	76%	75%
Logistic Regression	68%	60%
Random Forests, Decision Tree	50%	50 %

Figure 3. Predicting Match Results

6.2. Predicting Player Scores

The results for second problem statement were consistent across different algorithms. Therefore, we only chose Random Forests as our final algorithm.

We got different accuracies for different players of Pakistan since each player has different values of scores for each match and the patterns are also different for all players. One thing to note in this problem is that since this is a 3 class classification problem, accuracy of 70-80 percent is very good because a random guess would give 33 percent only. Looking at the accuracies of for different individual players, we get.

Player Name	Accuracy	Precision
Sharjeel Khan	71%	75%
Younis Khan	81%	88%
Kamran Akmal	80%	76%
Sarfaraz Ahmed	82%	81%
Shoaib Malik	74%	70%

Figure 4. Predicting Individual Player Scores

6.3. Predicting Next Wicket

The results for this problem were most interesting. Our accuracy came out to be 92 percent with 60 percent precision. We applied the trained model on a cricket match from 2007 which was not included in the train data. Our model successfully predicted 8 wickets out of 16 which is a great success rate considering the randomness in cricket. Moreover, the model gave 0 label for 120 values while the ground truth was 127 values. These results showed that applying neural networks to such a problem can find great results.

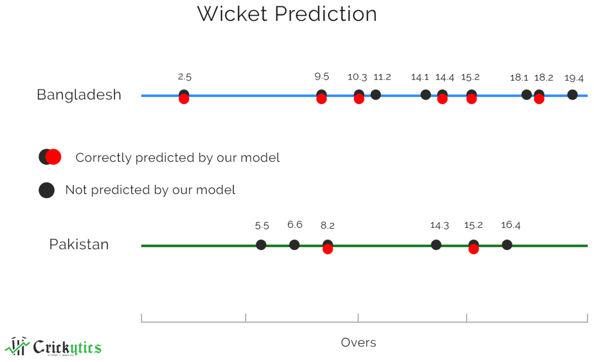

Figure 5. Wicket Prediction

7. CONCLUSION AND RECOMMENDATIONS

We successfully applied machine learning methods to problems of cricket prediction including predicting a match results, player scores and next wicket. We've seen that machine learning methods perform way better than random guessing. There is little research being done in the field and we hope that more people will start applying machine learning to cricket and draw great findings from their models.

Although machine learning approaches work very well in the context of cricket predictions, there are some fundamental problems that we had to face. The data for cricket matches is not enough for machine learning models to perform very accurate. A lot of feature engineering has to be done in order to make something good. Our future directions include predicting the results of tournaments and matches when they are happening.

REFERENCES

[1] Akash Malhotra, A Statistical Analysis of Bowling Performance in Cricket, 2015

[2] Stylianos Kampakis, Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches, 2015

[3] MG Jhawar, Predicting the Outcome of ODI Cricket Matches, 2016