Cricket game analysis

Shreya Singh, Saksham bairathi, Harsh Gauttam, and Aishwaryaditya Jha

This is project report for the subject ’CS1110: AI and ML’ at JK Lakshmipat University, Jaipur

ABSTRACT

Cricket is a second most popular sport played by about 120 million players. In this project we intended to predict winner and score of a cricket match and analyse the performance of a player in the upcoming match using different ML models. We executed our project in Python programming language. This report aims at solving the problem of predicting the results and score of game as well as predicting the cricket team by identifying the significant attributes from the data set and using the ML algorithms on them. We have limited our area of study to the World cup 2019 and IPL (2008-2020) game and 2015 ODI match(for score prediction). In cricket team prediction, we used different ML models to predict the best players for the playing eleven . In cricket team chase prediction , we used different ML models to predict the chasing probability considering the current run rate and number of wickets fallen. In ODI cricket match score prediction, we use random forest classifier method to predict score considering various probability like runrate.

Introduction

Cricket is one of the most popular and well-known outdoor game played in England, India and Australia on national and international level. According to ICC, cricket has over one billion fans (in the surveyed age category of 16-69) globally, the average age of which is 34 with a demographic breakdown of 61 percent male and 39 percent female. Fans were also asked about cricket and the Olympic Games with 87 percent of fans claiming they would like to see Twenty20 cricket in the Olympic Games. Also, most of the work that has been done on cricket forecasting is mostly relate to pre-match forecasting. Thus,demand for during-play forecasting is in high demand. There is a large amount of profit made in cricket betting even if you’re not an expert in the game as cricket is a game dominated by statistics. Studying historical results of the stadium and researching variables that have affected variables in the past will help in the prediction.The models that we present in our project, can serve various other purposes like analysis of team and player performance, identifying the key moments in a match, in addition to betting. We use the Logistic Regression Model, and Random Forest methods to build a forecasting tool that will predict if a team is going to win or lose by testing the current pre-match covariates and during-match statistics on the trained model. These forecasts of probabilities of win or lose can be done at any point as the match progresses.

For this project we tried to predict the winner and score of a match along with cricket team prediction and chasing probablity of cricket by using Python Programming and ML Modelling.
There were several challenges that we faced during the project as listed below:

- The data that we get from scraping, is a very raw data containing each and every information or attributes of all leagues of cricket. So the first challenge was to sort the data according to our need of analysis.
- After sorting, there were some empty values which we had to correct manually by selecting those particular rows and modifying as per our needs.
- Other challenges included scraping data from scrapped data multiple times and manually checking for any exception cases.

Background

History of Game of Cricket

Cricket was originated from south England. It is believed that the game is being played since 16th century. As the time passed the game caught the attention of many other countries and soon became the most played game in the world. It has been named the national game for the countries like England and Australia. There have been some evolution in the game like initially the bat used to play was shaped like a baseball bat but as the time passed and the game was played more and more people found it a bit uncomfortable playing with that bat and hence as the solution to the problem new shape was given to the bat, the same we have now. Also, the balls were properly shaped and made from leather.

Literature Review

[**Daniel Mago Vistro**, **Faizan Rasheed**, **Leo Gertrude David**] This paper aims to predicts the winner of the IPL match from season 2008 to 2017.For analysis, they have used several data science stages like pre-processing of data, visualization of data, feature selection, data preparation,and implementing different ML models which gives us different accuracy and thus helps us select the best model for prediction of the match outcome. Mainly they have used - Decision Tree model ,XGBoost classifier and Random forest classifier .Decision Tree model which gives good accuracy of 94.87 percent.XGBoost classifier which gives accuracy of 94.23 percent.Random forest classifier which gives accuracy of 80.76 percent ,which is quite less as compared to other models.

[**Rameshwari A. Lokhande and Pramila M. Chawan**]This paper aims to predict the score of first innings of the match on the basis of current run rate , number of wickets fallen during the inning, venue of the match and batting team. Second aim was to predicts the outcome of the match in the second innings considering the above attributes as mentioned along with the target given to the batting team. These two methods have been implemented using Linear Regression Classifier and Naïve Bayes Classifier. In both methods,they have made a 5 over intervals from 50 overs of the match and at each interval above mentioned attributes have been recorded of all non-curtailed matches played between 2002 and 2014 of every team independently.

[**Chetan Kapadiya**, **Ankit Shah**, **Kinjal Adhvaryu**, **Pratik Barot**] This paper aims for cricket team selection by predicting individual players performance using machine learning in different O.D.I, T20 International and test matches.In the analysis part, they have used many steps like pre-processing of data, visualization of data, feature selection, data preparation, and implementing different ML models which gives us different accuracy and thus helps us select the best model for prediction of playing eleven. Mainly they have used - Naive bayes algorithm with 58.12 percent accuracy, Decision trees with 86.50 percent accuracy and random forest algorithm with 92.25 percent accuracy but Weighted Random Forest algorithm gave the most accuracy with 93.73 percent.

[**Michael Bailey and Stephen R. Clarke**] This paper aims to predict the match outcome in on going O.D.I. cricket matches.They used the match data collected from all 2200 ODI matches played before January 2005. They have created a set of attributes for explaining statistically significant portions of variation associated with the predicted total runs and match results. They have used variables that include ground or turf advantage, past performances, match experience, performance at the precise venue, performance against the precise opposition, experience at the precise venue and current form. They used a multiple linear regression model for the research. Prediction variables were numerically weighted consistent with statistical significance and wont to predict the match outcome.

.

Methodology

We have followed the following methodology in our project. The methodology consists of 5 different stages as shown i.e. Problem Description, Data Description,Data pre-processing, Modelling and Evaluation of Results

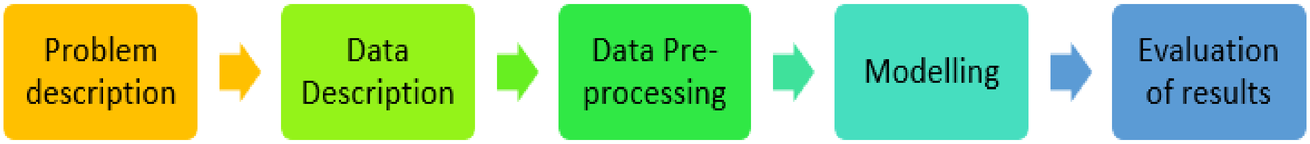


Figure 1. Methodology

0.1 Problem Description

Our project is divided is different objectives as follows:

1. Our First aim is to predict the winner of the matches of World Cup 2019 using previous three world cup matches results.
2. Our Second aim is to predict the score of cricket ODI match.
3. Our Third aim is to predict the playing eleven based on their performances in previous cricket matches.
4. Our Fourth aim is to predict the chasing probability of a team in a cricket match.

0.2 Data Collection and Data Cleaning

0.2.1 Winner prediction

We have used the scrap data containing 25 columns and 1952 rows. There are entries related to all leagues of cricket with attributes like Match id ,Team1 ,Team2, Gender ,Date, Season ,Series/Competition, Match number, Venue ,City ,Neutral venue, Toss winner, Toss decision ,Player of match ,Umpire1 ,Umpire2 ,Reserve umpire, Tv umpire, Match referee, Winner, Winner runs, Winner wickets, Method ,Outcome

1. **Results** The dataset spans across all the seasons of the world cup from 2007 to 2015 season. The data contains the match-wise winners of last 3 world cups.The data columns of this data contains the team wise information like Team1, Team2 and the winner. The data has 3 no. of columns and 128 no. of rows The data types present in this data is object. However, we had to tackle the missing values data. There were 2 matches whose 'winner' was not declared .So, I had to manually fill in the data of those 2 missing values for greater accuracy
2. **Ranking** The data contains the position of the team in world cup 2019.The data columns of this data contains the team wise position information like position, team name and points. The data has 3 no. of columns and 11 no. of rows The data types present in this data is int and object. The data obtained from was already cleaned. So, I did not have to do any sort of cleaning on the data.
3. **Fixtures** The data contains fixtures of the teams participating in world cup 2019.The data columns of this data contains the datewise match information like Round Number,Date, Location,Team1,Team2, Group. The data has 6 no. of columns and 49 no. of rows The data types present in this data is int and object. The data obtained from was already cleaned. So, I did not have to do any sort of cleaning on the data.
4. **worldcup 2019** The data contains team-wise information like previous appereances, previous titles, previous finals and semifinals record and current ranking of the team in world cup 2019 which is required for modelling the project. The data has 7 no. of columns and 11 no. of rows The data types present in this data is int and object. The data obtained from was already cleaned. So, I did not have to do any sort of cleaning on the data.

Score Prediction

1. **ODI match** The data contains the ball by ball information played by bowler batsman in an ODI match.The data contains the date, venue, batsman, bowler,runs per ball ,wickets fallen , date ,venue ,batting team, bowling team, striker ,non-striker and The data has 15 no of columns and 606 no of rows The data present in this data is int and object. The data obtained from was already cleaned. So, I did not have to do any sort of cleaning on the data.
2. **ODI** The data contains the ball by ball information played by bowler batsman in an ODI match.The data contains the date, venue, batsman, bowler,runs per ball ,wickets fallen , date ,venue ,batting team, bowling team, striker ,non-striker and The data has 16 no of columns and 308 no of rows The data present in this data is int and object. The data obtained from was already cleaned. So, I did not have to do any sort of cleaning on the data.

Cricket Team Predictor

1. **Batsmen** The data contains the record of batsmen performance since 2008 in the Indian Premier League (I.P.L.).The data set consists of 1300 rows and 16 columns. The columns of this data contains the player performance information consisting of the number of matches played and the average strike rate for a batsman. Now we had to tackle the missing values or "NaN" values in our data sets. Therefore we used the "dropna()" function in python.
2. **Bowler** The data contains the record of bowler's performance since 2008 in the Indian Premier League (I.P.L.).The data set consists of 700 rows and 13 columns. The columns of this data contains the player performance information consisting of the number of wickets taken and the economy of a bowler. Now we had to tackle the missing values or "NaN" values in our data sets. Therefore we used the "dropna()" function in python.
3. **All-Rounder** The data contains the record of the all-rounder's performance since 2008 in the Indian Premier League (I.P.L.).The data set consists of 1073 rows and 12 columns. The columns of this data contains the player performance information consisting of the number of sixers and wickets taken by the player. Now we had to tackle the missing values or "NaN" values in our data sets. Therefore we used the "dropna()" function in python.

Chasing probability of a team

1. **IPL** The data contains the record of the teams which are chasing in the Indian Premier League (I.P.L.).The data set consists of 8500 rows and 16 columns. The columns of this data contains the current batsmen, bowler and current run rate information . Now we had to tackle the missing values or "NaN" values in our data sets. Therefore we used the "dropna()" function in python.

0.3 Modelling

In this stage we will decide the model best suited for our prediction/objective. There are mainly five ML model approaches for cricket prediction.

1. Logistic Regression
2. Random Forest Classifier
3. Decision Tree algorithm
4. Gaussian support vector
5. Linear Regression

0.3.1 Winner Prediction

We will use Random Forest Classifier for this prediction. The Random Forest algorithm is an ensemble classification algorithm which create many decision trees during the training phase and each decision tree will predict different outcomes and then the algorithm predicts the classifier for all those decision trees and hence it gives the better accuracy rate from all the decision trees. To avoid over fitting and under fitting of the data is that we should include all significant variables in our data.We used stepwise approach of applying random forest classifier for a better explanation of the objective .

0.3.2 Score Prediction

We Use Linear Regression for this prediction. Linear Regression is a step up after correlation method and is used when we have to calculate one variable based on another variable.In this algorithm we model the relationship between dependent variables and one or more independent variables linearly.It tries to fit the variables in linear equations.It gives a better explanation of the objective

0.3.3 Cricket team predictor

We will use Gaussian support vector machine for this prediction. The Gaussian kernel computed with a support vector is an exponentially decaying function in the input feature space The SVM classifier with the Gaussian kernel is simply a weighted linear combination of the kernel function computed between a data point and each of the support vectors. Gaussian kernels are universal kernels i.e. their use with appropriate regularization guarantees a globally optimal predictor which minimizes both the estimation and approximation errors of a classifier. Prediction is done based on the 3 independent entities i.e. matches played, runs scored and strike rate.

0.3.4 Chasing probability of a team

We will use train, test and split to divide our data in two parts for training and testing our model for this prediction. Then we used logistic regression to test the accuracy of the model respectively. Prediction is done based on the 3 independent entities i.e. current run rate, runs scored and fall of wickets.

0.4 Procedure

0.4.1 Winner Prediction

1. Importing the required libraries like pandas,numpy ,train test split ,RandomForestClassifier
2. Import the csv file containing the results of matches played between 2006 and 2017
3. Displaying the match details of the teams playing in 2019 world cup and checking the teams defined above is only present in the past results dataset
4. Checking for correlation values on World cup 2019 details
5. For checking accuracy,define ’X’ variable containing whole dataset except winner column and define ’y’ variable containing only winner column.
6. Checking for any missing values in csv files
7. Dividing the training and testing dataset in 70 percent and 30 percent and calculating their accuracy separately using random forest classifier
8. Making the prediction dataset based on the current ranking of a particular team

9. Importing Fixtures of world cup 2019 and based on the created prediction set,winner of each match is predicted and displayed

10. Define a ’predict’ function which will take input of playing teams,ranking , past results and the applied classifier.

11. Define semi-finals and finals teams and with the use of ’predict’ function, get the results.

0.4.2 Score Prediction

1. Importing the required libraries like pandas, train test split, linear regression

2. Importing the csv file containing the results of an ODI match,odi between two teams.

3. The dataset contains ball by ball coverage of ODI match.

4. Dividing the training a testing dataset in 70 percent and 30 percent

5. Scaling the data and Fitting the data in Linear regression

6. Calculating the accuracy using Linear regression

7. Predicting the score pf Second batting team

0.4.3 Cricket team predictor

1. Importing the required libraries like pandas, numpy, Logistic Regression, sklearn.

2. Import the csv file containing the performance of the player since 2008 in IPL.

3. Displaying the player performance details playing in IPL .

4. For checking accuracy,define ’X’ variable containing whole dataset except player name column and define ’y’ variable containing only player name column.

5. Making the prediction dataset based on the past performances for players.

6. Took input of the number of batsmen, bowlers and all rounders from the user .

7. Added the predicted team into a data frame.

0.4.4 Chasing probability of a team

1. Importing the required libraries like pandas, numpy, Logistic Regression, sklearn.

2. Import the csv file containing the performance of teams since 2008 in IPL while chasing.

3. Displaying the team performance details while chasing in IPL .

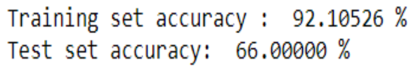
4. Dividing the training and testing dataset in 75 percent and 25 percent and calculating their accuracy separately using logistic regression.

5. Making the prediction dataset based on the past performances of the teams.

0.5 Evaluation of Results

0.5.1 Winner Prediction

Divided the data in 70 percent training dataset and 30 percent testing dataset and then the data gives the accuracy as shown below:



```
Training set accuracy :  92.10526 %
Test set accuracy:  66.00000 %
```

Figure 2. Accuracy of training and testing data

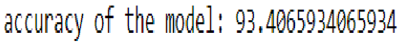
Exception in results

There is one exception occurred due to past results comparision which reduced accuracy of model and that exception is:

- In the semi-finals prediction, according to past results, India will win against new zealand but in reality new zealand won.

0.5.2 Score Prediction

Divided the data into percent training dataset and percent testing dataset and then the data gives the accuracy as shown below:

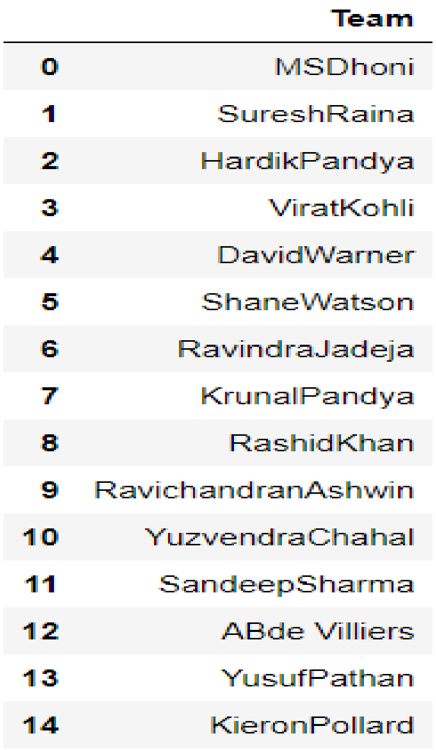


```
accuracy of the model: 93.4065934065934
```

Figure 3. Accuracy of the model

0.5.3 Cricket team predictor

Our classifier method was successful in creating a team which consisted of the players who had best performances over the few years.

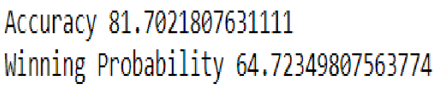


	Team
0	MSDhoni
1	SureshRaina
2	HardikPandya
3	ViratKohli
4	DavidWarner
5	ShaneWatson
6	RavindraJadeja
7	KrunalPandya
8	RashidKhan
9	RavichandranAshwin
10	YuzvendraChahal
11	SandeepSharma
12	ABde Villiers
13	YusufPathan
14	KieronPollard

Figure 4. Final output of the team predicted by the model

0.5.4 Chasing probability of a team

Our logistic regression method was successful in giving the chasing probability of a team .



```
Accuracy 81.7021807631111
Winning Probability 64.72349807563774
```

Figure 5. Accuracy of training and testing data and Winning Probability

Conclusion

Our first goal is to predict the winner of the world cup 2019 matches and to predict the cricket team.For winner prediction, we have used the data of three previous played world cup matches in order to design our model.We have use random forest classifier to design our model which gives greater accuracy as compared to other models.Random forest classifier uses multiple decision tree approach to predict the outcome which increases the efficiency of the model.

Our second objective was to predict the score of an ODI match which was done with the help of linear regression which gives more efficiency as compared to other models.

Our third goal is to predict a team based on the previous performances of the players in previous played IPL matches in order to design our model. We used Gaussian support vector machine for this prediction.

Our fourth goal is to predict the chasing probability of a team based on the performances of the teams in IPL match in order to design our model. We used rain, test and split to divide our data in two parts for training and testing. We divided he training and testing dataset in 75 percent and 25 percent and calculated their accuracy separately using logistic regression.

Future Scope

1. GUI can be included in the project for a better visualization.
2. The project can be made more generalised so as it can be used for different matches without making much changes in the base code.
3. More precision can be introduced in generation of the results by using different ML models.
4. These classification techniques can be used in analysing other sports like basketball,football etc.

Refrences

1. Daniel Mago Vistro, Faizan Rasheed, Leo Gertrude David- INTERNATIONAL JOURNAL OF SCIENTIFIC TECHNOLOGY RESEARCH VOLUME 8, ISSUE 09, SEPTEMBER 2019 ISSN 2277-8616
2. Rameshwari A. Lokhande-Student and Pramila M. Chawan- Professor(Computer and IT Dept, Veermata Jeejabai Technological Institute, Mumbai, India)-Prediction of Live Cricket Score and Winning-International Journal of Trend in Research and Development, Volume 5(4), ISSN: 2394-9333 www.ijtrd.com
3. Chetan Kapadiya,Ankit Shah,Kinjal Adhvaryu,Pratik Barot, Intelligent Cricket Team Selection by Predicting Individual Players' Performance using Efficient Machine Learning Technique, International Journal of Engineering and Advanced Technology (IJEAT) ,ISSN: 2249 – 8958, Volume-9 Issue-3, February 2020
4. Michael Bailey and Stephen R. Clarke, Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress,The 8th Australasian Conference on Mathematics and Computers in Sport, 3-5 July 2006, Queensland, Australia

Author's Contribution

- **Project Idea/Problem Formulation:** Shreya,Saksham,Aishwayaditya and Harsh
- **Data gathering** :Shreya,Saksham,Aishwayaditya and Harsh
- **Data Pre-processing:** Shreya,Saksham,Aishwayaditya and harsh
- **Literature review:** Shreya,Saksham,Aishwayaditya and harsh
- **Implementation :** Shreya -Winner prediction, Harsh - Score Prediction ,Saksham and Aishwayaditya -Cricket team prediction, Saksham and Aishwayaditya - Chasing probablity of a team
- **Result preparation and interpretation:** Shreya,Saksham,Aishwayaditya and Harsh
- **Report writing:**All sections- Shreya,Saksham,Aishwayaditya and Harsh
- **Poster designing:**Shreya