International T-20 Cricket Match Winner Prediction with Machine Learning

Intisar Tahmid Naheen (ITN), Md. Mominul Haque, Nahid Hossain Jibon

Abstract:

T20 is a popular form of cricket worldwide. The T20 style cricket is where players from different countries giving an explosive display of their performance in just 20 overs. Because of such a dynamic display of playing cricket, the T20s audience is growing day by day. And because of such popularity predicting T20 matches outcome has a profitable aspect to it. It can also be helpful for the cricket boards and people who bet on matches. In this paper, we have tried to predict the winning team of international t-20 matches based on the toss winner, toss decision, venue, and the city of the venue, teams, and which team is the home team. We used the decision tree model, random forest model, and other models to try to predict the winner using the available dataset from https://cricsheet.org/ It has the data from 2005-02-17 to 2021-03-20. The highest prediction accuracy is about 55%. Though we achieved a poor accuracy rate we learned a lot about machine learning and data analysis. We came up with the conclusion that significant facts to decide the winner prediction of a T20 match are team_1', toss_decision, team_2', 'city', 'toss_winner', 'venue', home_team. Based on these analyses, our proposed model determines the winner of sample matches from the data.

The code is available at:
https://github.com/nahidosen/Group_5_CSE445_3_Spring_21

1. Introduction:

SPORTS statistical analysis use in sports has been growing quickly year by year. Due to which the ways in which game strategies are formed or the player's evaluation criteria have been changed but also has got the more interest of the audience towards cricket. Now Cricket has become one of the most followed team games in the world with billions of fans all across the globe. Cricket is a sports game that is played globally across 106-member states of the International Cricket Council (ICC), which has 1.5 billion worldwide fans according to ICC. However, much of the global finance and interest is focused upon the 10 full ICC member nations and more specifically upon 'the big three' of England, Australia, and India. Cricket has evolved over time. Today, there are three major formats in which cricket is being played internationally, One Day Internationals (ODIs) and the T20 cricket and Test Matches. Besides these international cricket matches, T20 cricket is getting attention from the fans due to its shortest format and the most exciting format of the game.

Every year several countries that have earned the eligibility of playing T20 participate in different tournaments. Each national team consists of 11 players. Every team's performance is based on the key performances of players, team conditions, and other important aspects which decide the team's performances in a cricket match. The model will be built on all the possible factors affecting the outcome of a cricket match. Ground impacts, team quality, and home-field advantage were observed to be essential as well. This might be on the grounds that the ICC rating assesses results (win, draw, misfortune) alongside the success edge, wickets, and adversary rating. Winning the hurl was likewise considered in the model fitting however was observed to be insignificant. The playing conditions differ from ground to ground and nation to nation. For instance, playing conditions in Wankhede at Mumbai is very not quite the same as in Leeds at Headingley. Pitch

Conditions are very important in a cricket game. There are several kinds of pitches on which cricket has been played. Every ground and his own pitch conditions known for bowling pitches or batting pitches. To reset the target in interrupted matches, there is an approach using the name Duckworth-Lewis or D/L method. This winning prediction can be obtained using multiple classification methods depending on how much accuracy those offers while predicting the result.

2. Related works:

Factors to Anticipate Cricket winner

As IPL and t20 international cricket are quite alike we also have looked into research papers working on winner predictions of IPL. Abdullah Umar Nasib, Ajmain Inqiad Alam, and Mahfuzur Rahman in 2019, mentioned in their research paper, A Technique to predict Indian Premier League Match Winner using Artificial Intelligence, about the features that they had considered. Which were team1, team2, toss_winner, toss_ decision, city, venue, winner.

Cricket Winner Prediction Models

Daniel Mago Vistro, Faizan Rasheed, Leo Gertrude David in their paper, The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics used models such as Decision tree classifier, random forest classifier, XGBoost classifier.

Decision Tree Classifier

The Decision Tree algorithm belongs to the family of supervised learning algorithms. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior training data. For predicting a class label for a record it starts from the root of the tree. It compares the values of the root attribute with the record's attribute. On the basis of comparison, it follows the branch corresponding to that value and jumps to the next node.

Random Forest Classifier

Random forest is a flexible, easy-to-use, machine learning algorithm that produces great results most of the time even without hyper-parameter tuning. It is also one of the most used algorithms, because of its simplicity and diversity. Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest thus builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Accuracy Score

Daniel Mago Vistro, Faizan Rasheed, Leo Gertrude David in their paper, The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics mentions that they got the highest accuracy of 94.87% with the decision tree classifier.

3. Methodology

Our Research methodology is outlined in the figure below

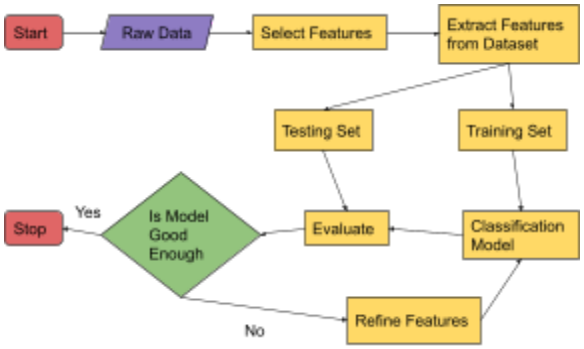


Figure: Applied Methodology During The study

A. Data Collection

The initial data was collected from Cricksheet. They provide ball by ball data for every format of

cricket. The data were in .yaml format first and was not the way we were expected the dataset would be. So, we extracted the necessary features from those datasets and made our own dataset compatible with the study.

B. Data Processing

The dataset that we came up with finally had 21 features in total. Among them, there were several features such as the win-loss ratio of both team1 and team2 which contained null values. Initially, we dropped those features which contained even a single null value to prepare our dataset to be efficient. After that stage, the number of remaining features was 10 in total which were dates, venue, city, home_team, team_1, team_2, toss_winner, toss_decision, umpire_1, Umpire_2, and match_winner. We used an encoded dictionary to encode the columns containing teams such as team_1, team_2, home_team, toss_winner & match_winner. Other than that we used label encoder venue, city, and toss_decision. After applying classifiers, the model performed well in the training dataset but performed low in the test dataset. To improve the performance we decided to drop dates, umpire_1 and umpire_2 from the data frame and found a slightly better accuracy rate.

C. Data Description

Feature Name	Description	Data Type
dates	When the match hosted	object
venue	Venue of the match	object
city	Where the venue is	object
home_team	The city belongs to the team	object
team_1	Team 1	object
team_2	Opposition team	object
toss_winner	The team wins the toss	object
toss_decision	The winner decides to bat or bowl	object
umpire_1	First umpire	object
umpire_2	Second umpire	object
match_winner	The team won the match	object

D. Classification

The dataset is shuffled and then split into 75% training and 25% testing sets. After that, different machine learning algorithms, including Decision Tree, Random Forests and Support Vector Machines were applied to find performance in predicting the acceptance of the papers.

Decision Tree: Visualizing and understanding a decision tree is easy, and we found that an entropy-based decision tree with seven nodes deep is optimum for us. The decision tree works with information gain and entropy to split the features such that the data is distributed in the tree as heterogeneously as possible. Eq. (5) shows the formula for entropy, which is calculated for every feature, and the highest is selected as the root node:

$$E(s) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

where S is the subset of training examples and pi is the probability of the class.

Random Forest: Random forest consists of many decision trees contributing to the prediction that is the modal class. Random forest classifier turns out to be the best classifier,

with 10,000 estimators using the entropy method. Since using 10,000 estimators was very resource-heavy, we went with 1,000 estimators as the accuracy

Support Vector Machine: We used a support vector machine with a Radial Basis Function (RBF) kernel as the optimum classifier. Support vector machines work by constructing one or more hyperplanes that can best separate the data.

4. RESULTS

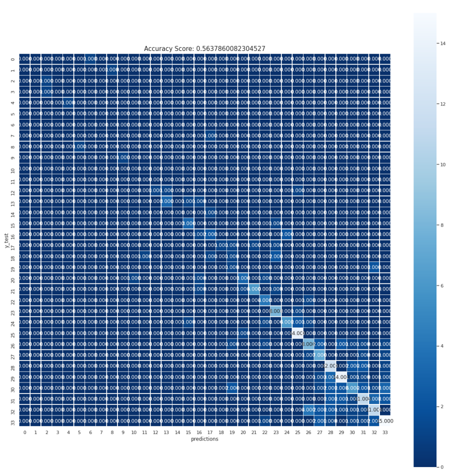
We applied 3 machine learning models and got the highest accuracy of 56.37%.

Decision Tree Classifier

We split the test and train data to X and y and the cross-validation scores for 4 iterations were:

Cross-Validation Score	0.4526749	0.43209877	0.46090535	0.40909091
Train Accuracy	0.9629120879120879			
Test Accuracy	0.5637860082304527			

Confusion Matrix



Measures of Performance

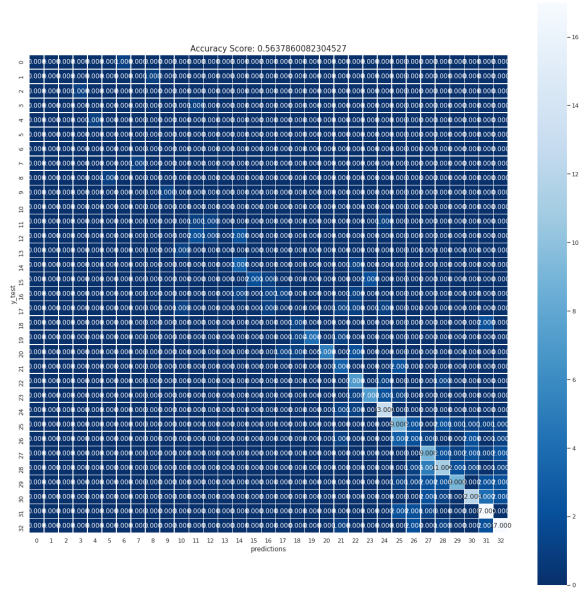
	precision	recall	f1-score	support
9	0.00	0.00	0.00	1
10	0.00	0.00	0.00	1
13	0.50	1.00	0.67	1
14	0.00	0.00	0.00	1
15	1.00	1.00	1.00	1
17	0.00	0.00	0.00	0
18	0.00	0.00	0.00	0
20	0.00	0.00	0.00	1
21	0.00	0.00	0.00	1
22	1.00	1.00	1.00	1
26	0.00	0.00	0.00	0
28	0.00	0.00	0.00	0
29	1.00	0.33	0.50	3
30	0.75	0.60	0.67	5
31	0.00	0.00	0.00	1
32	0.60	0.75	0.67	4
33	0.25	0.20	0.22	5
34	0.00	0.00	0.00	4
35	0.00	0.00	0.00	5
36	0.17	0.33	0.22	3
37	0.60	0.50	0.55	6
38	0.86	0.67	0.75	9
39	0.50	0.80	0.62	5
40	0.62	0.89	0.73	9
41	0.75	0.55	0.63	11
54	0.78	0.93	0.85	15
58	0.53	0.47	0.50	17
59	0.47	0.70	0.56	10
70	0.55	0.71	0.62	17
71	0.67	0.67	0.67	21
72	0.46	0.32	0.37	19
77	0.55	0.52	0.54	21
90	0.58	0.50	0.54	22
98	0.62	0.65	0.64	23
accuracy			0.56	243
macro avg	0.41	0.41	0.40	243
weighted avg	0.56	0.56	0.55	243

Random Forest Classifier

Cross-validation scores for 3 iterations were:

Cross-Validation Score	0.49074074	0.55555556	0.49845201
Train Accuracy	0.9615384615384616		
Test Accuracy	0.5390946502057613		

Confusion Matrix



Measures of Performance

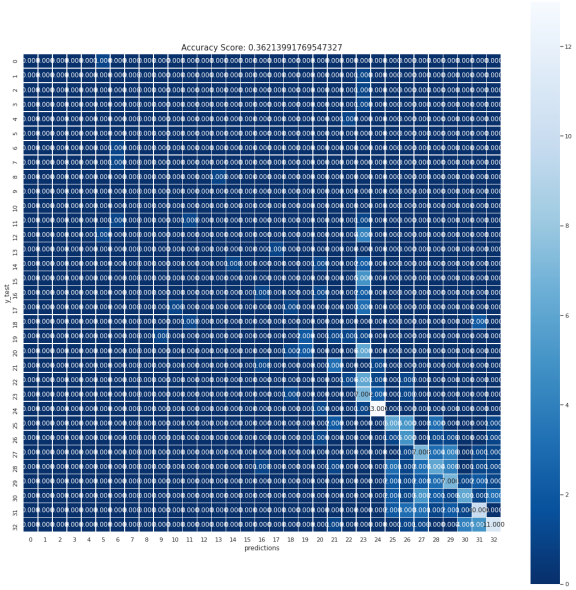
	precision	recall	f1-score	support
9	0.00	0.00	0.00	1
10	0.00	0.00	0.00	1
13	0.00	0.00	0.00	1
14	0.00	0.00	0.00	1
15	1.00	1.00	1.00	1
17	0.00	0.00	0.00	0
18	0.00	0.00	0.00	0
20	1.00	1.00	1.00	1
21	0.00	0.00	0.00	1
22	1.00	1.00	1.00	1
28	0.00	0.00	0.00	0
29	0.25	0.33	0.29	3
30	0.50	0.20	0.29	5
31	0.00	0.00	0.00	1
32	0.50	0.75	0.60	4
33	1.00	0.40	0.57	5
34	0.33	0.25	0.29	4
35	0.00	0.00	0.00	5
36	0.33	0.33	0.33	3
37	1.00	0.67	0.80	6
38	1.00	0.56	0.71	9
39	0.38	0.60	0.46	5
40	0.50	0.78	0.61	9
41	0.78	0.64	0.70	11
54	0.72	0.87	0.79	15
58	0.50	0.53	0.51	17
59	0.20	0.20	0.20	10
70	0.45	0.53	0.49	17
71	0.55	0.52	0.54	21
72	0.69	0.47	0.56	19
77	0.67	0.57	0.62	21
90	0.59	0.77	0.67	22
98	0.65	0.74	0.69	23
accuracy			0.56	243
macro avg	0.44	0.42	0.42	243
weighted avg	0.58	0.56	0.56	243

SVM

Cross-validation scores for 3 iterations were:

Cross-Validation Score	0.29012346	0.32407407	0.26006192
Train Accuracy	0.7486263736263736		
Test Accuracy	0.36213991769547327		

Confusion Matrix



Measures of Performance

	precision	recall	f1-score	support
9	0.00	0.00	0.00	1
10	0.00	0.00	0.00	1
13	0.00	0.00	0.00	1
14	0.00	0.00	0.00	1
15	0.00	0.00	0.00	1
18	0.00	0.00	0.00	0
20	0.33	1.00	0.50	1
21	0.00	0.00	0.00	1
22	0.00	0.00	0.00	1
26	0.00	0.00	0.00	0
28	0.00	0.00	0.00	0
29	0.50	0.33	0.40	3
30	0.00	0.00	0.00	5
31	0.00	0.00	0.00	1
32	1.00	0.25	0.40	4
33	0.00	0.00	0.00	5
34	0.33	0.25	0.29	4
35	0.00	0.00	0.00	5
36	0.00	0.00	0.00	3
37	0.50	0.33	0.40	6
38	0.00	0.00	0.00	9
39	0.38	0.60	0.46	5
40	0.33	0.11	0.17	9
41	0.17	0.64	0.27	11
54	0.76	0.87	0.81	15
58	0.38	0.35	0.36	17
59	0.25	0.50	0.33	10
70	0.33	0.41	0.37	17
71	0.32	0.29	0.30	21
72	0.37	0.37	0.37	19
77	0.50	0.29	0.36	21
90	0.45	0.45	0.45	22
98	0.58	0.48	0.52	23
accuracy			0.36	243
macro avg	0.23	0.23	0.21	243
weighted avg	0.37	0.36	0.35	243

Summary

Classifier	Training Accuracy	Testing Accuracy
Decision Tree Classifier	0.962912087 9120879	0.563786008 2304527
Random Forest Classifier	0.961538461 5384616	0.539094650 2057613
SVM	0.748626373 6263736	0.362139917 69547327

5. Conclusion and future works:

In this paper, we present machine learning approaches to predict the winning team of an international T20 match. As of now, we could not find any proper machine learning approaches in this kind of area. There are several papers published emphasizing the world cup winner or run prediction and some people worked on IPL T20 league winner prediction. So, we could not find any reference related to our area. We believe we are the ones who initiated this first and we have to admit that our model did not perform much well on the test dataset. However, the training accuracy was somewhere close to 96% which is pretty impressive. The most accuracy we could come up with on test data was 56% by applying both the Decision Tree and the Random Forest classifier.
Therefore, we are planning on collecting some more data in the future and try to come up with a better performance. Other than that, we have another plan to work on forecasting the total score on this international T20 format for each of the playing teams.

References:

https://cricsheet.org/

Daniel Mago Vistro, Faizan Rasheed, Leo Gertrude David, 2019, The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics

http://www.ijstr.org/final-print/sep2019/The-Cricket-Winner-Prediction-With-Application-Of-Machine-Learning-And-Data-Analytics

Abdullah Umar Nasib, Ajmain Inqiad Alam, and Mahfuzur Rahman, 2019, A Technique to predict Indian Premier League Match Winner using Artificial Intelligence

https://www.researchgate.net/publication/332254928_A_Technique_to_Predict_Indian_Premier_League_Match_Winner_using_Artificial_Intelligence