Predicting the winner of T20 world cup 2021 using Machine learning

Salman Farshi
Id : 1520162042
Salman.farshi@northsouth.edu

ABSTRACT

We love cricket and the most loving cricket format is T20. T20 world cup is knocking the door in India and we are eagerly waiting for the event. Every cricket lover wants to know the winning team of the tournament but this cannot be possible before plying the game. In this case we can use machine learning for advance predating. We have used different types of machine learning algorithm for getting better result. We achieved 64% accuracy in logistic Regression that is very effective for t20 matches. Other's machine Learning algorithm used: Random Forest, KNN, Support vector machine. Our prediction is tested properly and we found a good outcome for prediction in different case.

Keyword: Winner, T20 cricket, Machine learning

Introduction:

Cricket is the most played game in the Indian subcontinent. Most the of country's cricket fan are crazy for cricket. Icc decided that Icc T20 world cup will be held in India in 2021 and that brings Indian subcontinent's people a joy of happiness. India is one of the best crickets played team in the world and they also won icc t20 world cup as well. India is the best known for their cricket achievement they played very well in the home town. Others team like Pakistan, Bangladesh, Srilanka , Afghanistan , England also played very good cricket in the Indian field   .

Machine leaning is the field of data science and the uses of it increasing day by day in every spare of life including cricket. We have observed a lot implantation of machine learning in cricket like predicting the runs, winning percentage etc.. For past motivation we want to Predict the match result as well winner of Icc T20 world cup of 2021 . In this paper we have showed different types of winning possibility using different types of machine learning.

Related works:

I reviewed a lot of articles and papers regarding cricket winning prediction. Shashank Singh [1] showed in his papers some important features and achieved a good accuracy. Shashank Singh used a special feature called Features Score where he added some with respect to each column's values. He got highest accuracy on random forest classifiers which is about 73% that is amazing accuracy for t20 cricket. Abdul Basit [2] showed in his paper some custom method which was very helpful for getting a good accuracy. He got highest accuracy in random forest classifier; the accuracy was about 64.68%. Surprisingly he archived 80.6 % accuracy on same random forest classifier but in this time, he used some custom mythology where $ustom\ Accuracy = Actual\ Value - Predicted\ Value$. Others papers also showed some important features and technique those were very helpful this paper.

Methodology

For this paper I collected data from different types of website such as ESPN cricinfo , Icc official website, crick sheet etc. Data was not such a good condition that I could fit the model at that time. I have spent huge amount time just to modify our data. I have changed data features for many times. Finally, I found some useful correlation based on ranking then I finalize my data set.

Date	Team_1	Team_2	Winner	Margin	Ground
Feb 17, 2005	New Zealand	Australia	Australia	44 runs	Auckland
Jun 13, 2005	England	Australia	England	100 runs	Southampton
Oct 21, 2005	South Africa	New Zealand	New Zealand	5 wickets	Johannesburg
Jan 9, 2006	Australia	South Africa	Australia	95 runs	Brisbane
Feb 16, 2006	New Zealand	West Indies	tied	NaN	Auckland
...	...	...	...	...	...
Feb 26, 2020	South Africa	Australia	Australia	97 runs	Cape Town
Feb 27, 2020	Kuwait	U.A.E.	U.A.E.	102 runs	Al Amerat
Feb 29, 2020	Thailand	Singapore	Singapore	43 runs	Bangkok

Result	Team1	Team2	Country	Team_ranking	Batting_ranking	Bowling_ranking	Winning_rate
1	England	Australia	England	1.5	3.0	1.0	50.0
0	England	Sri Lanka	England	1.5	3.0	1.0	50.0
0	England	Pakistan	England	1.5	3.0	1.0	50.0
0	Australia	England	England	1.5	3.0	1.0	50.0
0	England	West Indies	England	1.5	3.0	1.0	50.0
...	...	...	...	...	...	...	...
1	Afghanistan	West Indies	Afghanistan	0.0	0.0	3.0	58.0
1	Afghanistan	West Indies	Afghanistan	0.0	0.0	3.0	58.0
1	Afghanistan	Ireland	Afghanistan	0.0	0.0	3.0	58.0
1	Afghanistan	Ireland	Afghanistan	0.0	0.0	3.0	58.0
0	Afghanistan	Ireland	Afghanistan	0.0	0.0	3.0	58.0

Model: Machine learning process start with data collection and after data collection data prepossessing phase start. Only a well prepossessed data can be fit into the model . Sometimes we need to modify our model according to different prediction. Here is the Machine learning workflow .
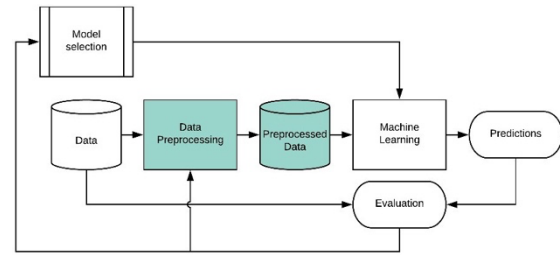
In this paper I have used four different classifiers 1. SVC ,2Ranom Foresr,3Logistic Regression, 4 KNN. For SVC , I firstly tried the base model of SVC , after getting output I modified this base model using others paremeter like

'C':[0.001,0.01,0.1,0.5,1],'gamma':['scale','auto']

```
cv='warn',
error_score='raise-deprecating'
estimator=SVC(C=1.0, cache_size=200, class_
weight='balanced',coef0=0.0, decision_funct
ion_shape='ovr', degree=3,'auto_deprecated'
, kernel='rbf', max_iter=-1,probability=Fal
se, random_state=None, shrinking=True, tol=
0.001, verbose=False), iid='warn', n_jobs=N
one,
 param_grid={'C': [0.001, 0.01, 0.1, 0.5, 1
],'gamma': ['scale', 'auto']},pre_dispatch=
'2*n_jobs', refit=True, return_train_score=
False ,scoring=None, verbose=0
```

Modified verson of SVC perform better than base verson . Random forest I start with base model and found some satisfactory output but when I tried to modify the model then I found that modified version did not work properly that is why it gave some less accuracy. The parameter for random forest are :

```
'n_estimators': 610,
 'min_samples_split': 4,
 'min_samples_leaf': 15,
 'max_depth': 3
```

Logistic Regression is the best model for my data set. Logistic model with best gave the best accuracy. Modified version with parameter are:

```
'solver': 'liblinear', 'C': 78.47
```

Based model of KNN perform does not well, it gives some unexpected accuracy then I fine-tuned the model and found better accuracy , the best parameter for knn is K=20 .

Result: Among mentioned model the best accuracy was found in logistic regression which is 64% . 2nd highest accuracy is 62% from SVC .

Without best parameter		With best parameter	
1 Support vector classifier	62%	1 Support vector classifier	62%
2 Random forest classifers	59 %	2 Random forest classifers	58 %
3.kKNeighborsClassifier	58%	3.kKNeighborsClassifier	63%
4 logistric Regressor	62%	4 logistric Regressor	64%

Maximum KNN score on the test data: 61.91%
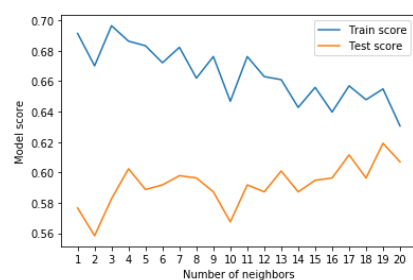
Base Model :

```
127]: {'Logistic Regression': 0.6282245827010622,
       'KNN': 0.5811836115326252,
       'Random Forest': 0.5993930197268589}
```

Conclusion and Future work:

I want to increase the accuracy level and also add some features if needed. This model should be tested in real data as I cannot test properly due to short time . Others model without SVC, KNN , logistic regression and random forest classifier should be implanted so that some better accuracy could be found .

References

1.Quantitative Analysis of Forthcoming ICC Men's T20 World Cup 2020 Winner Prediction using Machine Learning

2. ICC T20 Cricket World Cup 2020 Winner Prediction Using Machine Learning Techniques

[3] K. Passi and N. Pandey, "Increased Prediction Accuracy in the Game of Cricket Using Machine Learning," Int. J. Data Min. Knowl. Manag. Process, vol. 8, no. 2, pp. 19–36, 2018, doi: 10.5121/ijdkp.2018.8203.

[4] M. Yasir, L. I. Chen, S. A. Shah, K. Akbar, and M. U. Sarwar, "Ongoing Match Prediction in T20 International," vol. 17, no. 11, pp. 176–181, 2017.

[5] C. S. Abhishek, K. V Patil, P. Yuktha, K. S. Meghana, and M. V Sudhamani, "Predictive Analysis of IPL Match Winner using Machine Learning Techniques," Int. J. Innov. Technol. Explor. Eng., vol. 9, no. 2S, pp. 430–435, 2019, doi: 10.35940/ijitee.b1043.1292s19.

[6] G. S. K. Paramesha, "A Perspective on Analyzing IPL Match Results using Machine Learning," vol. 7, no. 03, pp. 1476–1479, 2019.

[7] V. Q. L. N. Hdvhu et al., "Player ' s Performance Prediction in ODI Cricket Using Machine Learning Algorithms."

[8] P. The et al., "Predicting The Cricket Match Outcome Using Crowd Opinions On Social Networks: A Comparative Study Of Machine Learning Methods. pp 63-76," vol. 30, no. 1, pp. 63–76, 2017.

[9] M. G. Jhanwar and V. Pudi, "Predicting the outcome of ODI cricket matches: A team composition based approach," CEUR Workshop Proc., vol. 1842, no. September, 2016.

[10] K. Desai, "Predicting Outcome of ODI Cricket Games," vol. 3, no. 01, pp. 1–5, 2015.

[11] S. Kumar and S. Roy, "Score Prediction and Player Classification Model in the Game of Cricket Using Machine Learning," vol. 9, no. 8, pp. 237–242, 2018.

[12] D. M. Vistro, F. Rasheed, and L. G. David, "The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics," vol. 8, no. 09, pp. 985–990, 2019.

[13] A. Nimmagadda, N. V. Kalyan, M. Venkatesh, N. Naga, S. Teja, and C. G. Raju, "Cricket score and winning prediction using data mining," vol. 3, pp. 299–302, 2018.

[14] S. Agrawal, S. P. Singh, and J. K. Sharma, "Predicting results of Indian premier league T-20 matches using machine learning," Proc. - 2018 8th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2018, pp. 67–71, 2018, doi: 10.1109/CSNT.2018.8820235.

[15] J. Kumar, R. Kumar, and P. Kumar, "Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks," ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun., pp. 343–347, 2018, doi: 10.1109/ICSCCC.2018.8703301.

[16] I. Technology and S. Ms, "a Survey on Team Selection in Game of Cricket Using Machine Learning," pp. 2442–2446, 2019.

[17] M. B. Alvi, N. A. Mahoto, M. A. Unar, and M. A. Shaikh, "An Effective Framework for Tweet Level Sentiment Classification using Recursive Text Pre-Processing Approach," no. June, 2019, doi: 10.14569/IJACSA.2019.0100674.

[18] M. B. Alvi, N. A. Mahoto, M. Alvi, M. A. Unar, and M. A. Shaikh, "Hybrid Classification Model for Twitter Data - A Recursive Preprocessing Approach," no. April, 2018, doi: 10.1109/IMTIC.2018.8467221.

[19] Ghani, M. U., Alam, T. M., Jaskani, F. H. (2019, November). Comparison of Classification Models for Early Prediction of Breast Cancer. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-6). IEEE.