Odi Cricket Prediciton

Aditya Yadav
adiyadav.1729@gmail.com

Aniket Dhawad
aniketdhawadiitg1@gmail.com

Vikash Goyal
vikashgoyal2701@gmail.com

September 3, 2019

Abstract

The aim of this study was to investigate to what degree a machine learning model can capture outcome of a One Day International (ODI) match depending on team's composition and strength. The target competition was the World Cup 2019. Basic features comprising of batting and bowling metrics was used along with features capturing the impact and quality of players. Team's batting and bowling strength was created based on playing squad for a match and venue where the match was played. The model was tested with different training dataset based on timeline. Result showed how greatly ODI cricket has changed with time.

1 Introduction

One Day International(ODI) Cricket is one of the three formats which are played at International Level. International Cricket Council(ICC), governing body of cricket, has currently 12 Full Team Members and 92 Associate Members. ODI cricket started in 1971 as second form of cricket after test cricket to cater entertainment to larger set of spectators.In this format, both teams are allowed to play for fixed overs and the one with more runs is declared winner. Originally started with 60 overs and white jerseys, ODI cricket has gone through various transition over years to stay relevant and consistently entertaining. 50 overs , powerplay overs, field restrictions, batting friendly grounds are among few significant changes that has played role in its evolution.

Cricket World Cup, the most important tournament in this format generally played every 4 years, with preliminary qualification rounds leading up to a finals tournament. The tournament is one of the world's most viewed sporting events and is considered the "flagship event of the international cricket calendar" by the ICC. World cup 2019 has 10 teams participating. In league round all 10 teams would be playing against each other. Top 4 would march into semi- finals. Winner of the semi-finalist would be playing for the coveted title. There would be 45 league matches, 2 semi-finals and 1 finals. This paper aimed at predicting the result of 48 matches.

We collected data from cricinfo and relianceiccranking. Over 100 features were created which included players and teams features as well. Pearson correlation, chi square test and recursive feature elimination were used for feature elimination. The selected features were used as inputs to three different classification algorithms: logistic regression, random forest, boosted decision trees. Ensemble of weak classifiers were also tested. Best performing model was selected as final one to predict the outcome.

2 Methodolgy

2.1 Data

Data was scrapped from www.espncricinfo.com. Firstly, list of all matches played was scrapped. Secondly, list of all players played for different country was scrapped. Note that data had match_id, player_id which was key in getting different information at match and player level respectively. We used www.relianceiccrankings.com to collect information of player batting and bowling ranking at the start of every month from 1971 till date.

Therefore, for every match we had batting and bowling scorecard, player details at match level and their ranking.

Wicket Order	Wicket Weight
1	1.5
2	1.4
3	1.5
4	1.3
5	1.2
6	0.9
7	0.7
8	0.5
9	0.5
10	0.3
11	0.2

Table 1: Wickets Order and Weightage .

2.2 Feature Engineering

2.2.1 Basic Metrics

Players Batting Average, Strike rate, Bowling Average , Bowling Economy, Bowling Strike Rate was considered just before a match for which features were generated.

To compare like to like players from two teams, players role was identified. That is, features were arranged such that openers from a team would be compared against openers from its opposition. Similarly, for bowlers , first change bowlers would be observed against similar role bowler from opposition team

2.2.2 Impact Factor

This factor was created to compare players based on its quality rather than amount of runs scored/wickets taken. Impact features take into account the quality of opposition bowlers against which runs scored. A batsman scoring runs against a bowler who has a bowling average of 25 will have higher impact score compare to scoring runs against a bowling average of 35.

Similarly, Bowlers taking higher order wickets are given higher weightage for batsmen dismissed. This could lead to a situation where bowler with a high average(or poor strike rate) can get higher weightage on wickets. Hence, strike rate was taken into factor.

$$ImpactFactor, Bowler = \frac{weightedsumofwickets}{wickets * strikerate}$$

Matt Henry till Feb 2019 had close to 45% top order wickets(1-3) out of 78 batsmen he dismissed. Lasith Malinga on other hand had close to 38% of top order wickets in his 322 wickets. Henry's took wickets at strike rate of 29 whereas Malinga's took struck on every 32 balls. Each of these features was taken to calculate their impact factor.

Players	Matches	Wkts	Weighted Wkts	SR	Impact
Matt Henry	43	78	91.2	28.7	1686
Lasith Malinga	218	322	329	32.6	990

Table 2: Impact Factor : Bowler.

2.2.3 Form Factor

Batting:
While comparing batsmen , generally batting average is used a metric to identify the batsmen in form. Though, it is reasonable to comparable batsmen on average when a considerable time period is taken into account. However, it does not give clear picture of batsmen form in last few(say 10) matches. For instance, a batsmen with score sequence 120,0,30,40,10 another with score 60,45,45,35 have batting average of 50. Though it is fair to assume second batsmen is more consistent.

Runs	Balls	Standard Deviation	Rec Avg * SR	Career Avg * SR	Form Factor
89,82*,2,10*,6	115,89,6,6,7	43	5340	3672	2.1
27,0,90,45,91	39,1,99,46,50	33	3861	4842	1.02

Table 3: Form Factor.

In addition to runs scored, rate at which they are scored also play a crucial role in fate of a match. This is also taken in consideration to determine batsmen's recent form.

For instance, before England's league match against India in world cup 2019, Ben Stokes and Bairstow had following streaks. Bairstow is more of a consistent batsmen(2nd row) with a standard deviation of 33 , however comparing Stokes recent number with overall number, he is actually punching above his weight and thus had a higher form factor.

Bowling:

This metric was calculated to capture the form of the bowler in the recent few matches before the actual match. A bowler's performance can be quantified using his bowling average, bowling strike rate and economy. Lower the value of these metrics for a bowler in a particular match, better the performance. So a new metric was created which basically took the harmonic mean of these three performance metrics and ultimately this harmonic mean was averaged out for each player as per the last 5 matches he played.

Rashid Khan was ranked as the 3rd best bowler in ODI cricket with a rating of 747 where as Mohammad Amir was ranked 36th with a rating of 544. But the recent form of Amir before the Afghanistan vs Pakistan match was way better than Rashid. This scenario was captured using this metric.

Mohammad Amir

Overs	**Runs**	**Wkts**	**Econ**	**Opposition**	**Start Date**
10	67	2	6.7	England	03-Jun-19
10	30	5	3	Australia	12-Jun-19
10	47	3	4.7	India	16-Jun-19
10	49	2	4.9	South Africa	23-Jun-19
10	67	1	6.7	New Zealand	26-Jun-19

Form Factor **0.85**

Rashid Khan

Overs	**Runs**	**Wkts**	**Econ**	**Opposition**	**Start Date**
DNB	-	-	-	New Zealand	08-Jun-19
7	45	0	6.42	South Africa	15-Jun-19
9	110	0	12.22	England	18-Jun-19
10	38	1	3.8	India	22-Jun-19
10	52	0	5.2	Bangladesh	24-Jun-19

Form_Factor **18.27**

2.2.4 Team Strength

For every team, batting and bowling strength was calculated by considering the ratings of players of the team as per the ICC world rankings. The batting strength was calculated by taking the average rating of top 6 batsmen and bowling strength by taking the average rating of top 5 bowlers. We assumed the minimum rating for player who was not present in the top 100 list.

Pitch Weighted Team Strength: In order to take into the impact of the pitch, we tried to quantify whether the pitch was a high scoring or a low scoring one. We did this by calculating the average runs that are scored on a particular pitch. If the pitch was high scoring one, we increased the batting strength of a team and reduce the bowling strength of a team by a particular factor.

Ground	Year	Pitch Impact			
Rajkot, 82*,2,10*,6	115,89,6,6,7	43	5340	3672	2.1
27,0,90,45,91	39,1,99,46,50	33	3861	4842	1.02

Table 4: Form Factor.

For Example, Rajkot has always been high scoring ground whereas low and slow wickets of Delhi has generally been difficult to score.

Rationale behind this feature was that high scoring pitches are reasonably flat and give advantage to batsmen and reduce the advantage to bowlers.

2.3 Modelling

2.3.1 Feature Selection

Over 120 features were created. Pearson correlation scores were calculated to identify and hence remove correlation among features. Features such as Batting average and strike rate was highly correlated and they were multiplied to create another feature. Similarly, bowling average, strike rate and economy was clubbed to make one feature. Also, features like team win rate was introducing bias and hence this was also not shortlisted in final list of feature.

2.3.2 Train and Test Data

One day cricket had undergone immense change since its inception. Hence selection of training data became an important factor. On experimenting with training data of different period , we identified post 2001 period as best suited training data for our model. Post 2017 one day matches were taken as test data.

2.3.3 Modelling

Finally 78 features were selected were model. We tried logistic regression , Extreme Gradient Boosting and Random Forest to generate predictions. Since, sports data is more likely to have noise, Random Forest is suitable as it is robust in dealing with noise. Extreme Gradient Boosting , owing to its characteristics of capturing even noise, was over-fitting and hence resulted in poor performance.

Random Forest gave an accuracy of 70% in test data.

Escape special TeX symbols (Compress whitespace

References

[1] www.cricinfo.com

[2] http://www.relianceiccrankings.com/

[3] Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches - Stylianos Kampakis, William Thomas

Date	Team1	Team2	Prediction	Probability	Actual
Thursday, May 30	England	South Africa	England	0.57	England, 104 runs
Friday, May 31	West Indies	Pakistan	Pakistan	0.635	West Indies, 7 wickets
Saturday, June 1	New Zealand	Sri Lanka	New Zealand	0.732	New Zealand, 10 wickets
Saturday, June 1	Afghanistan	Australia	Australia	0.52	Australia, 7 wickets
Sunday, June 2	South Africa	Bangladesh	South Africa	0.51	Bangladesh, 21 runs
Monday, June 3	England	Pakistan	England	0.63	Pakistan, 14 runs
Tuesday, June 4	Afghanistan	Sri Lanka	Afghanistan	0.61	Sri Lanka, 34 runs
Wednesday, June 5	South Africa	India	India	0.56	India, 6 wickets
Wednesday, June 5	Bangladesh	New Zealand	Bangladesh	0.58	New Zealand, 2 wickets
Thursday, June 6	Australia	West Indies	Australia	0.67	Australia, 15 runs
Saturday, June 8	England	Bangladesh	England	0.69	England, 106 runs
Saturday, June 8	Afghanistan	New Zealand	Afghanistan	0.54	New Zealand, 7 wickets
Sunday, June 9	India	Australia	Australia	0.59	India, 36 runs
Wednesday, June 12	Australia	Pakistan	Australia	0.59	Australia, 41 runs
Friday, June 14	England	West Indies	England	0.69	England, 8 wickets
Saturday, June 15	Sri Lanka	Australia	Australia	0.75	Australia, 87 runs
Saturday, June 15	South Africa	Afghanistan	South Africa	0.57	South Africa, 9 wickets
Sunday, June 16	India	Pakistan	India	0.61	India, 89 runs
Monday, June 17	West Indies	Bangladesh	Bangladesh	0.5779	Bangladesh, 7 wickets
Tuesday, June 18	England	Afghanistan	England	0.5339	England, 150 runs
Wednesday, June 19	New Zealand	South Africa	New Zealand	0.569	New Zealand, 4 wickets
Thursday, June 20	Australia	Bangladesh	Australia	0.6775	Australia, 48 runs
Friday, June 21	England	Sri Lanka	England	0.7287	Sri Lanka, 20 runs
Saturday, June 22	India	Afghanistan	India	0.5733	India, 11 runs
Saturday, June 22	West Indies	New Zealand	New Zealand	0.7121	New Zealand, 5 runs
Sunday, June 23	Pakistan	South Africa	Pakistan	0.5733	Pakistan, 49 runs
Monday, June 24	Bangladesh	Afghanistan	Bangladesh	0.5075	Bangladesh, 62 runs
Tuesday, June 25	England	Australia	England	0.674	Australia, 64 runs
Wednesday, June 26	New Zealand	Pakistan	New Zealand	0.6274	Pakistan, 6 wickets
Thursday, June 27	West Indies	India	India	0.7208	India, 125 runs
Friday, June 28	Sri Lanka	South Africa	South Africa	0.7617	South Africa, 9 wickets
Saturday, June 29	Pakistan	Afghanistan	Pakistan	0.5469	Pakistan, 3 wickets
Saturday, June 29	New Zealand	Australia	Australia	0.656	Australia, 86 runs
Sunday, June 30	England	India	England	0.6188	England, 31 runs
Monday, July 1	Sri Lanka	West Indies	Sri Lanka	0.5419	Sri Lanka, 23 runs
Tuesday, July 2	Bangladesh	India	Bangladesh	0.5022	India, 28 runs
Wednesday, July 3	England	New Zealand	England	0.6455	England, 119 runs
Thursday, July 4	Afghanistan	West Indies	Afghanistan	0.5163	West Indies, 23 runs
Friday, July 5	Pakistan	Bangladesh	Bangladesh	0.5461	Pakistan, 94 runs
Saturday, July 6	Sri Lanka	India	India	0.721	India, 7 wickets
Saturday, July 6	Australia	South Africa	Australia	0.6207	South Africa, 10 runs
Tuesday, July 9	India	New Zealand	India	0.62	New Zealand, 18 runs
Thursday, July 11	Australia	England	Australia	0.64	England, 8 wickets
Sunday, July 14	New Zealand	England	England	0.58	England, superover