CS584 – MACHINE LEARNING FALL 2016

MODERN CRICKET SIMULATION FOR TWENTY-TWENTY FORMAT

Table of Contents

Modern Cricket Simulation for Twenty-Twenty format

Task

The task is to classify the outcome of the next ball in a Twenty-twenty(T20) cricket match as one of the below mentioned categories.

- Boundaries [outcome=4 or 6]
- Runs[outcome=1,2,3]
- Wickets [Batsman got out]
- Defensed[outcome=0]
- Extras

Dataset

The Dataset used in the project consists of 530 csv files where each file represents a single match data. Each match file consists of 240 balls data.

Data source

Our primary data source is cricsheet.com. We have collected additional data from espn cricinfo for bowler and batsman details. Few columns were created from the available data.

Target variable

There is a single target variable Output. The output consists of 5 values.

Features

The input features are the batting details, bowling details, bowler and batsman performance in the venue. There are 14 final input features.

Data size

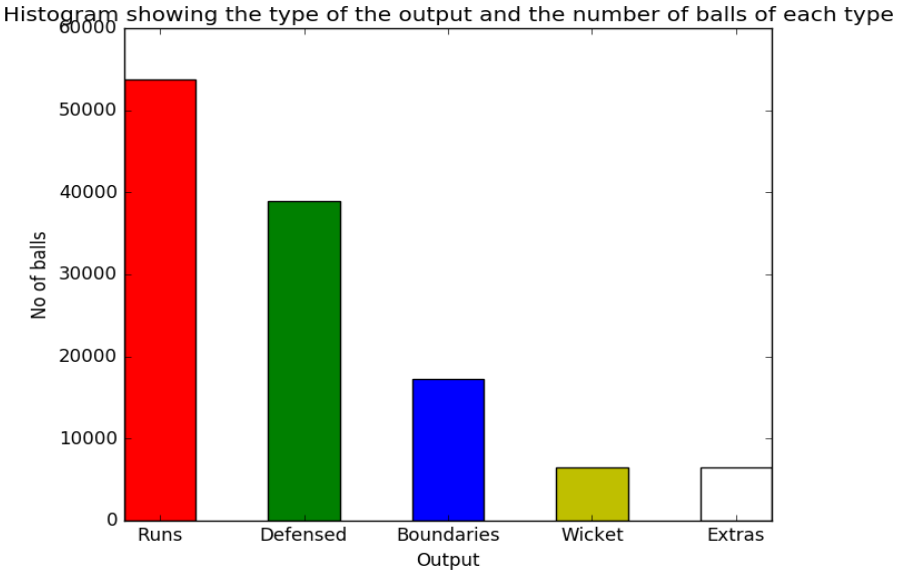
There are 123062 instances.

Preprocessing

- Data from cricsheet.com and espn cricinfo are joined.
- Few columns which are not required in the analysis are removed.
- Few new columns were generated from the available data.
- All the features are scaled using sklearn preprocessing package.

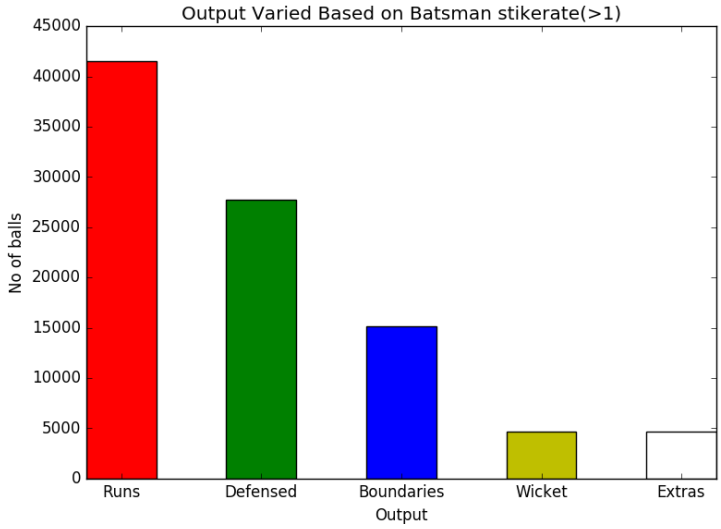
Visualization

Target

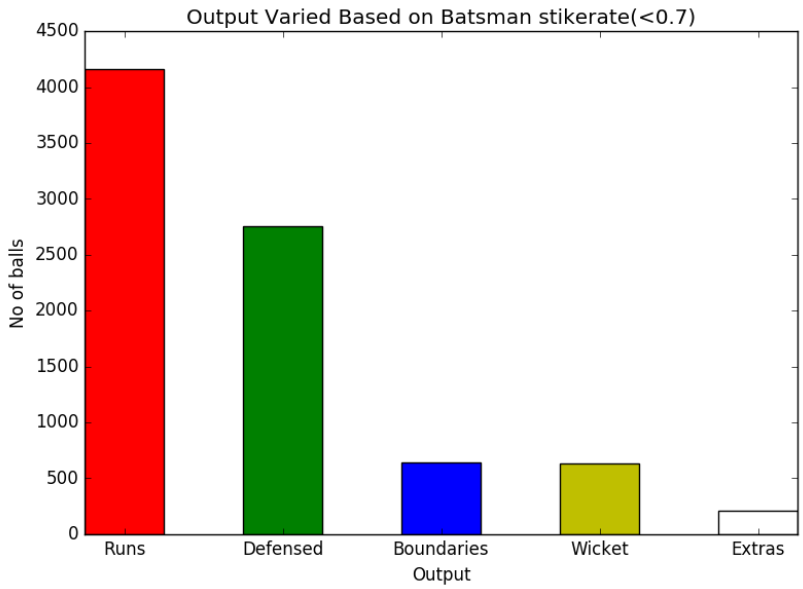


The above histogram shows how the output is classified over the full data. It is clear that wicket and Extras are the minimal occurred events whereas Runs is the most probable event.

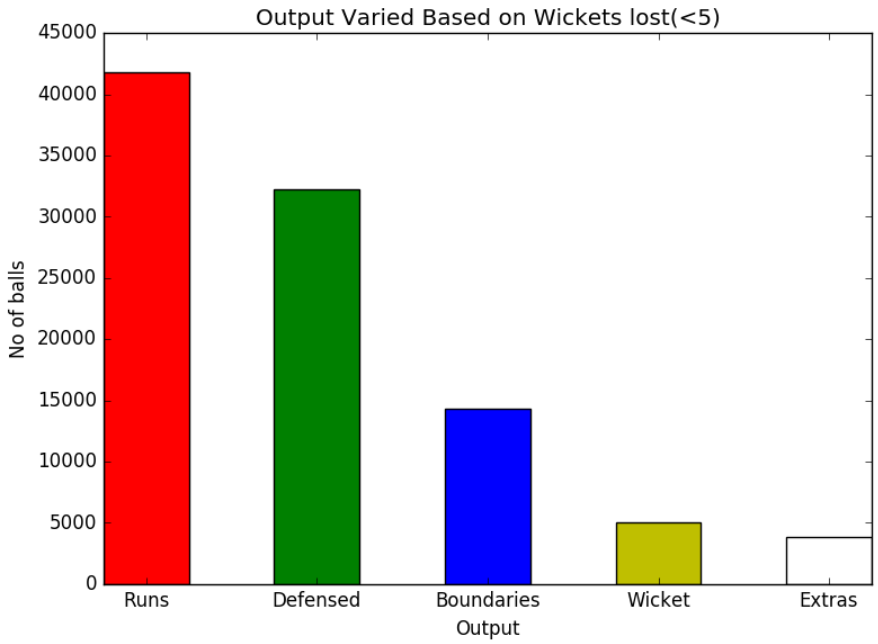
Features



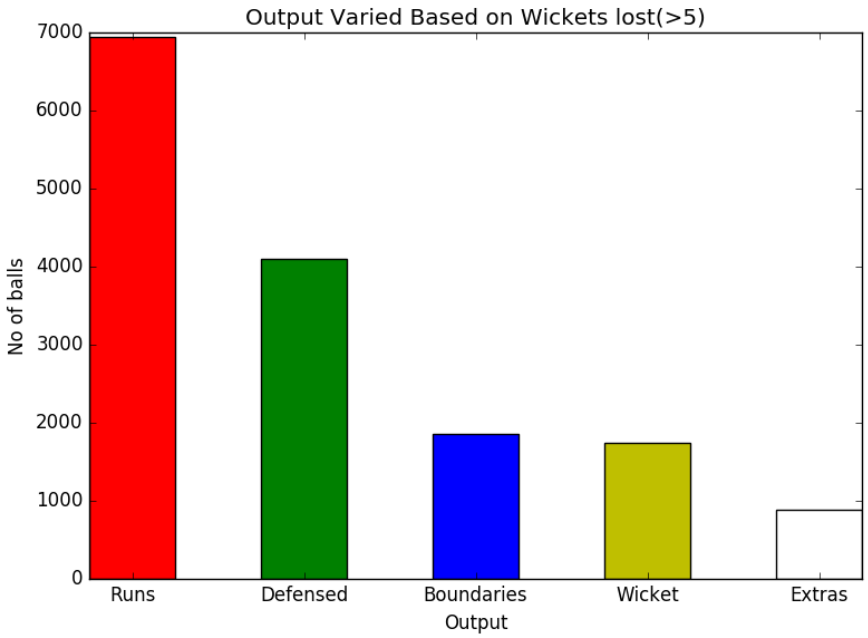
As we can see from the histogram it is clear that when the strike rate of the batsman is greater than 1(when batsman is able to score more runs compared to the number of balls) then the number of boundaries are very high. If we can observe the probability of losing the wicket is also higher.



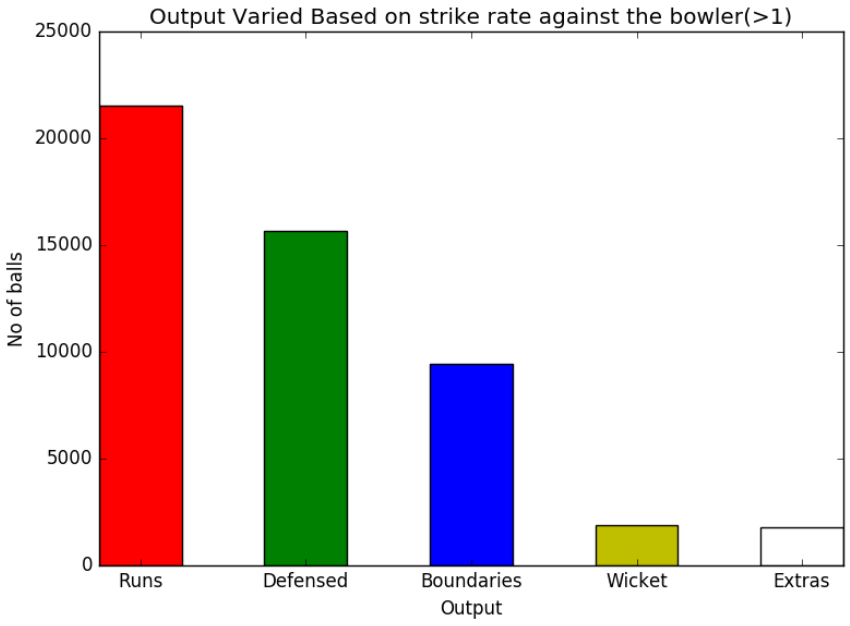
It clear that when the batsman is scoring at a lesser rate than the number of balls, he is in a defensive mode of saving his wicket so he will not attempt to hit boundaries. The number of wickets lost is higher than the number of boundaries. The number of wickets lost is less when compared at a higher strikerate.



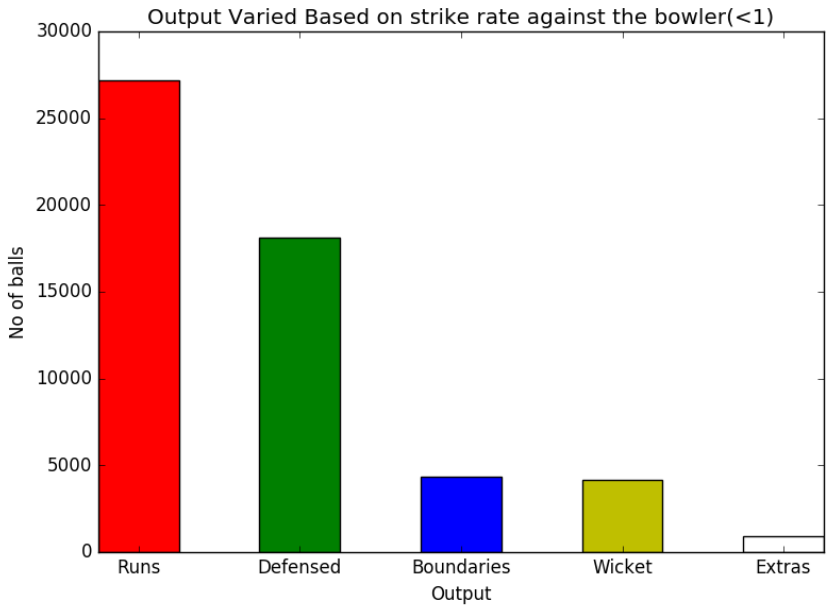
Number of wickets lost is also a major factor to consider. When the wickets lost is less batman tends to be aggressive and try to score more boundaries. Here the proportion of the boundaries is higher.



When the team lost 5 wickets they will be in a defensive state and try to score more Runs compared to boundaries. It is very clear from the histogram that the number of wickets lost is nearly equal to number of wickets lost.



When the strike rate of the batsman is higher for a particular bowler, he is assumed to score more Runs and Boundaries and the number of defensed will be less. The chance of losing the wicket is also less.



When the strike rate of the batsman is less for a particular bowler then the number of Runs and Defensed will be high and the number of wickets lost will be high.

Evaluation

Performance Measure

We choose accuracy as a measure. We are unable to calculate other measures using scikit learn.

Classifiers

We have used Naïve Bayes, Multinomial Naïve Bayes, multinomial Logistic Regression with Newton-cg solver and sag solver. We have tried these classifier as the problem we are trying to solve is a multilevel classification.

Evaluation Strategy

We have tried cross validation as we felt that the data is very less to perform test train split. We used 10-fold cross validation.

Performance Results

Model	Parameters	Performance
DecisionTreeClassifier	CV= 5	0.548089581377

	CV=10	0.469990975714
Logistic regression	solver ='newton-cg',multi_class='multinomial‘,	0.610269774222
	solver ='sag',multi_class='multinomial	0.610269774222
Multinomial NB	alpha=0.1	0.450130229583
	alpha = 1	0.450097733037
GaussianNB	CV=10	0.562001387025

Top Features

- All Strike rates (Bowler and batsman)
- Number of wickets lost
- Batsman performance with venue

Discussion

The best classifier is the logistic regression. The best classifier was unable to give proper results as the available data is not enough to classify the output into multi levels. We don’t have proper data to classify an extra. The data is not enough to properly classify the Defensed state. Logistic regression performed well when compared to all other classifiers.

Interesting/Unexpected Results

- Normally, we found a situation where the strike rate of the bowler is high and the strike rate of the batsman is low and the batsman performance in this venue is low
- In this situation we expect the coming ball to be a defended or runs and it is runs but the model predicted it as a boundary.
- There are many situations like this in our results.

Conclusion

We are able to get a 61% accurate results. We feel that the data is not enough to make proper analysis. We need some more details about the ball to clearly classify the outputs. We are in the plan of working on this project even after the class. We should have considered the extras as outliers as we don’t have a proper data to classify the ball as extras and even the number of occurrences is also less.

References

Data source: http://cricsheet.org/, http://www.espncricinfo.com/

https://pdfs.semanticscholar.org/4667/1ddcbb7bcee189ede56937c440b2ec4d0147.pdf

Notes (erase these notes before you save and submit):