Improving Duckworth lewis method by using machine learning

MSc Research Project
Data Analytics

Muthu Sam Jaipal
x16132483

School of Computing
National College of Ireland

Supervisor: Symeon Charalabides

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Muthu Sam Jaipal
Student ID:	x16132483
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Symeon Charalabides
Submission Due Date:	11/12/2017
Project Title:	Improving Duckworth lewis method by using machine learning
Word Count:	5803

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Improving Duckworth lewis method by using machine learning

Muthu Sam Jaipal
x16132483
MSc Research Project in Data Analytics

11th December 2017

Abstract

There has been lot of research has been done in the area of sports with prediction. Machine learning algorithm has been applied on most of the data to predict its outcome. In cricket also there are many research has been done to find out the outcome of the match. But there are a special case in the cricket where statistic plays an important role. When the cricket match is interrupted by the rain and then again it starts during the shortage of the time it is difficult to predict the run and overs remaining in the cricket match.To over come this issue Duck worth lewis method has been introduced. In this paper we are going to discuss the short come of the method of Duck worth lewis method. In this research We are going to improve the model of duck worth lewis method with machine learning models. At end we are going to compare the Duck worth lewis method, Duck worth lewis method with run rate as a variable and Improved version of duck worth lewis method.

1 Introduction

Cricket is a sport which is invented by England which is otherwise called as a gentlemen's game.Cricket is played in most of the sub-continent countries. These countries are India, Pakistan, Srilanka,Bangladesh with some European countries such as England, Scotland, and Ireland. Cricket is a game of uncertainties were to predict the outcome of match is very difficult. There is a emerging field in the computer science that is machine learning algorithm. Most of the time machine learning is used for prediction.And many author has used this for predicting the cricket match. As the cricket match are played in the open ground there can be chances of abrupt of the match due to rain, lights and wet outfields McHale and Asif (2013). When the match is adrubted due to rain it very difficult for management of the cricket to decide how much run and over is remaining. To solve this problem two English statistician frank Duck worth and Tony-lewis has proposed some mathematical calculation to predict the runs and over of the second team in the limited overs match.Later this method was adopted by the International cricket council to predict remaining over and runs. According to this method there are two variables has been chosen to calculate the remaining runs and overs McHale and Asif (2013).These two resources are Remaining bowl and Wicket. These two variables are given more importance According to Singh et al. (2015). But these are not only the variables which plays an important role in predicting the run and over remaining in the match. Some other factors such as run rate, Toss , Home team,Players rating,pitch etc are some of the variables which plays an important in predicting Overs and wickets. So in this thesis we have made some changes in the algorithm which calculates the run rate of the match.We are going to use the machine learning algorithm such as Neural network,Support vector machine,decision tree,logistic regression,binomial logistic regression to predict the overs and wicket. we are going to compare the model of duck worth lewis method, Duck worth lewis method with the Run rate and duck worth lewis method improved version. In this research we are going to see the performance of these methods with the machine learning algorithm

2 Related Work

Cricket is the second most popular sports after football. Many prediction has been done on different sports using Data mining algorithm . Sankaranarayanan et al. (2014) has done the prediction of the outcome of the match that is winning or losing . They have used the linear regression and nearest neighbour clustering algorithms to predict the outcome of the match. In this research author has used the historical match data to predict the outcome of one day match.But they have failed to conduct the test on the short term cricket match like T20 matches which is played only for 20 over. Bandulasiri (2008) In this paper author has analyzed the factor such as Winning the toss ,match type(Like day or night) because day or night plays an important role in the field of predicting the match output. In the day time bowls rotation is towards outward of the field and similarly in the night it is much difficult for the bowler to grip the bowl because of wet field. So it becomes very difficult for the bowling team to bowl in the night time which gives the advantage to the batting team. But in this paper author has failed to make a point on the run rate. Run rate is an important aspect of the match because run rate is the shows the performance of the team in a match .Run rate can be calculated per over or by Per innings and should be taken into consideration. The concept of resource is developed by the duck worth and lewis method to solve the problem of resource management which helps in the prediction of the run and over required for the second inning of the match. Beaudoin and Swartz (n.d.) In this author has used two variables such as best batsman and best bowler but it is very difficult to calculate the best batsman because it is not feasible to evaluate the performance metric of bowler and batsman. But there is a slightest chance if we can check the rating of the cricketer. We need to see that physical strength of the player can also plays an important role in the predicting the match according to Kaluarachchi and Aparna (2010) but they have used other factor for example Winning toss,day and night effect ,home game advantage these are the factor used buy using the artificial intelligent techniques such as Baysian classifiers with machine learning technique. In this research mathematical calculation has been done to to predict the outcome of the match. But they have not used an important factor such as the influence of the umpire in the match. Umpire plays an important role in the match because any wrong decision in the match can lead to the winning of the defending team.In this paper author has created a tool called Cric AI which predicts the match of upto four and five matches in a row. (McHale and Asif (2013)) Author has attempted to make some modification in the duck worth lewis mehod. This is the first attempt to make some changes in the algorithm of the duck worth lewis method.Here the modified Duck worth lewis method makes changes in the two factor they are Estimation of F(w) and the model of Z(u,w). This is a great approach to make the changes there is a lot of improvement has been made in this research for example it has smoothed the variable of z0 from 288.6(for D/ method subjectively smoothed(f(w)) to 291.9 (D/L method smoothed f(w)) to(Modified D/L Model,Smoothed f(w)) to 340.0. This is a great achievement in terms of the accuracy of the duck worth lewis method because it will calculate the resource and outcome of the match more accurately.

Singh et al. (2015) has considered 2 data mining techniques in there research they are Linear regression and Nave bayes algorithm for there research. Here they have calculated the algorithm with the five overs interval they are 1-5 overs of 50 overs for both the algorithm. The outcome shows that the accuracy percentage of the algorithm increases from 68 percentage initially from 1-5 overs to 91 percentage till they reach 45th over. This shows that as the data increases the output of the match increases proportionally the accuracy of the model increases.

Phanse and Deorah (2011) has attempted to overcome the shortcoming of the duckworth lewis method. Authors have used the weka tool to use the algorithm such as Random forest method and C4.5. Here author has found out some of the limitation of the duck worth lewis method. From there research using weka tool they have found out that Venue Is the one of the important factor to estimate or predict the match score for second innings. So the author has used over , wickets and venue as the attribute for there research in data mining and they have come to a interesting fact and a conclusion that team batting first has more advantage then the team batting second from there results. That is true because when the first team bats first it has no target set so it has less liable to have pressure of the match. like wise if we see the second team has more pressure as compare to the defending team because they have a target in front of them which they have to achieve within a given over. The main drawback of thie paper is the bias factor has not been taken as an important factor. The over all ratio or the the probability of the match win has been calculated by matches win and the number of the matches played by the team. The main conclusion of this research is as follows. 1) 66 percentage of the match won by the team who has won the toss 2) The winning ratio of the team batting first is 64 percentage 3) Team choose to bat second during the rain effected matches these are 54 percentage 4) more stress has been given to the Wickets 5) more stress has been give to the matches 6) And the last is the match effected ratio is 82 percentage

Ramanayaka et al. (2016) has proposed the theory of getting out Is related to or we can say as it is associated with the number of factor that is number of bowl, Number of wicket , venue, run rate, rating of the bowler, pitch condition. But in this research the author has considered two factors they are number of wickets and number of bowls remaining. This two variables is directly dependent on the duck worth lewis method.As we know that the probability of the a particular bowl bowled by any bowler can be(6,4,2,1,0,3) or it can be getting out. According to the result we can see that getting out is directly proportional to the number of bowls and number of wickets. As the number of wicket and number bowl increases there are probability of getting out and if there is less wicket and bowl there is a less chance of getting out of the match. Here in the research they have found out that in the beginning of the over and at the end there is a lot of chance of getting out as compare in the middle over. But still it all depends on the bowler , batsman and wicket as well

Xu (2012)In this paper we have can see that researcher has used two most important algorithm to test they are genetic algorithm and Artificial neural network. They are the first author to use these algorithm to predict the outcome of the cricket. They have collected the 1000 data among which they used 700 data for testing and 300 for training the model. The result of medium square error for predicted and the original score is between 26.637 and is 32.334 for the test sample values. This concludes that the predicted medium square error is less then that of original with the genetic and artificial neural network algorithm. There are different feature has been used by the author for there prediction they are fatigue ,weather ,experience, time training, height, weight, nutrition facts etc with different grades like short ,less ,Long ,Above, tall, excellent and week.

Pathak and Wadhwa (2016) Here the author has used the variables such as physical fitness of the team ,dynamic strategies of the match. Because different captains have different strategy for the match . Author has used the data mining techniques such as Nave baysian techniques,random forest and support vector machine to predict the outcome of the match. They have also used some of the other variables such as home game advantage. It is true that home team gets more support in terms of the spectators cheering there team.Day or night match in this kind of match second team who is playing under lights have many disadvantage like vision,wet outfield and more swing in the bowl later. These are some of the important variables which are used in the match. but the variable called the physical strength of the team cannot be calculated accurately as there are only 10 player in each team who plays in the field. Here user failed to take a variable called average winning rate of the team in the same ground and against same team because this is the metric which is very important to consider the winning team with there victory.

2.1 Machine learning used in other sports

Prediction of the sports is not only in restricted to cricket but it has also done in many other sports such as football score prediction. In the basketball such as number of baskets by team in the match. Me et al. (2011) has created a prediction model in the sports of swimming which is very interesting.In this research author has modelled there algorithm with fuzzy logic . Fuzzy logic is very powerful algorithm to tackle the continuous data. Because the data which is used in this research are continuous data. The data which is collected are from poland swiming club . The data which is collected are from the professional swimmers who have experince of atleast 7 years. The variables which are used in this research are age hours,Before-after,Time,Stimulus,land water,Sex.

Joseph et al. (2006) has predicted prediction football match . They have predicted the outcome of the match such as Win ,loss and draw.The data which has been selected from tottenham football club.The machine learning algorithm which is used is bayesian network . There are some other methods used such as MC4,Naive bayes,K nearest neighbour and Data driven Bayesian has been used When the model was built all the model has worked well but the performance of the K nearest neighbour was much better then the other algorithm.

There is another approach used by Martins et al. (2017) that is classifier. Here the author has used the polynomial algorithm to do the perdition of the outcome of the football match that is win, loss and draw. In this model 96 percentage of the accuracy has been achieved by the author. Here researcher has also compared the old works of the author with the accuracy of 0.52,0.68,0.92,0.53,0.56,0.58. etc

Hucaljuk and Rakipović (2011) main approach was to use the classifier.The main classifier which are used are naive bayes,Logit boost,random forest, and artificial intelligence.In order to predict the output of the football match author has built there own tool to predict it .Author has built in total two models they are basic and expert. Basic model has less accuracy as compare to that of the expert model. In this model random forest algorithm has shown more accuracy then the other model. In this paper author has failed to show the variables used for there research.

3 Methodology

After some research and analysis author has adopted the cross industry standard process for data mining (CRISP-DM). According to the pole conducted by KdNuggets the CRISP-DM is the most popular methodology. Here we have adhered all the phases for this research. We have used following phases Business understanding Data Understanding Data preparation Modelling Evaluation Deployment.

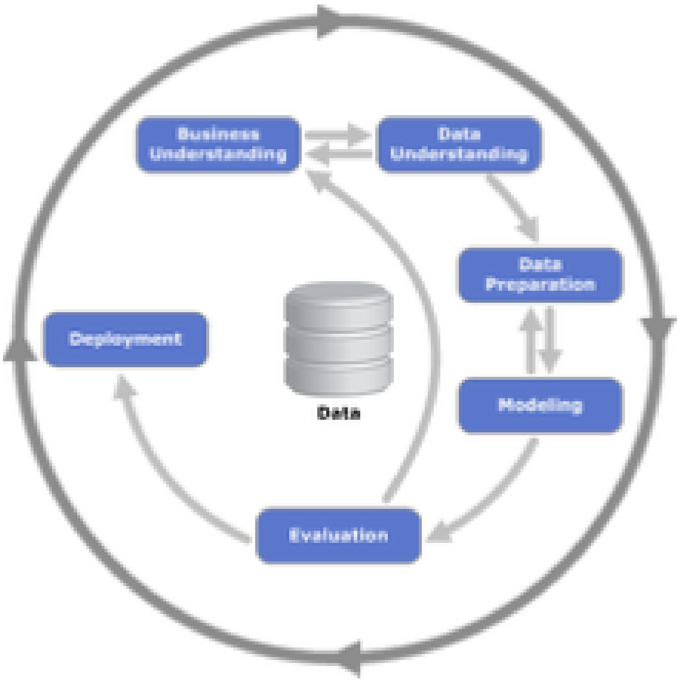


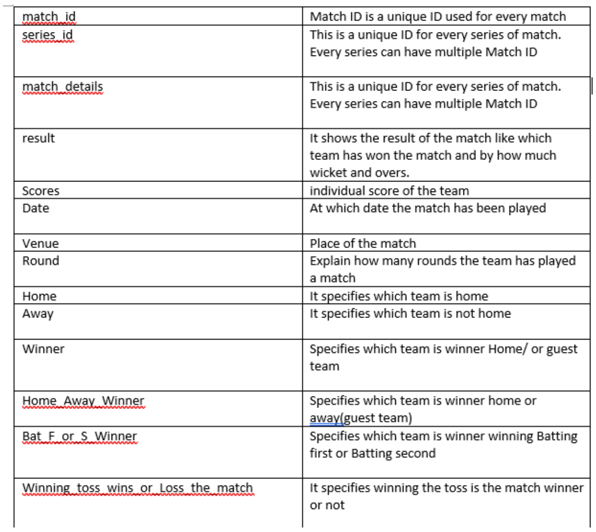
Figure 1: Cross industry data process

3.1 Data Processing

Data processing is the stage where we process the data to the desired format for our investigation on our data.Mainly there are three stages are included they are data collection, data cleaning and data conversion. Data processing is done in order to find out any noise or any unwanted data which can dilute our machine learning models

3.2 Data Collection

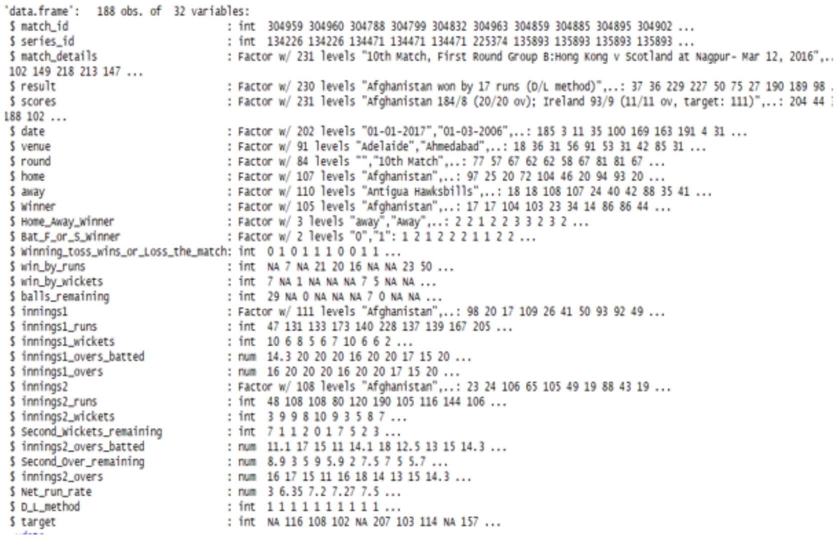
As the research is on cricket the data set has been taken from https://cricsheet.org/. which is the largest data set which contains the data of all the format of the matches such as one day cricket, Test and twenty-twenty format of the game. The data was in yaml format the data has been converted into excel by using converters .After conversion of the format we have the below data.



match_id	Match ID is a unique ID used for every match
series_id	This is a unique ID for every series of match. Every series can have multiple Match ID
match_details	This is a unique ID for every series of match. Every series can have multiple Match ID
result	It shows the result of the match like which team has won the match and by how much wicket and overs.
Scores	individual score of the team
Date	At which date the match has been played
Venue	Place of the match
Round	Explain how many rounds the team has played a match
Home	It specifies which team is home
Away	It specifies which team is not home
Winner	Specifies which team is winner Home/ or guest team
Home_Away_Winner	Specifies which team is winner home or away(guest team)
Bat_F_or_S_Winner	Specifies which team is winner winning Batting first or Batting second
Winning_toss_wins_or_Loss_the_match	It specifies winning the toss is the match winner or not

Figure 2: Cleaned data

After collection of the data its very important to clean the data.Data cleaning is the most important part of the data processing. This is because if we dont clean the data the effect of this will be seen till the end of our research. There has been an intensive cleaning has been done. Below is the imported data variable in the R studio.



```
'data.frame':   188 obs. of  32 variables:
 $ match_id                        : int  304959 304960 304788 304799 304832 304963 304859 304885 304895 304902 ...
 $ series_id                       : int  134226 134226 134471 134471 134471 225374 135893 135893 135893 135893 ...
 $ match_details                   : Factor w/ 231 levels "10th Match, First Round Group B:Hong Kong v Scotland at Nagpur- Mar 12, 2016",..
 102 149 218 213 147 ...
 $ result                          : Factor w/ 230 levels "Afghanistan won by 17 runs (D/L method)",..: 37 36 229 227 50 75 27 190 189 98 .
 $ scores                          : Factor w/ 231 levels "Afghanistan 184/8 (20/20 ov); Ireland 93/9 (11/11 ov, target: 111)",..: 204 44 :
188 102 ...
 $ date                            : Factor w/ 202 levels "01-01-2017","01-03-2006",..: 185 3 11 35 100 169 163 191 4 31 ...
 $ venue                           : Factor w/ 91 levels "Adelaide","Ahmedabad",..: 18 36 31 56 91 53 31 42 85 31 ...
 $ round                           : Factor w/ 84 levels "","10th Match",..: 77 57 67 62 62 58 67 81 81 67 ...
 $ home                            : Factor w/ 107 levels "Afghanistan",..: 97 25 20 72 104 46 20 94 93 20 ...
 $ away                            : Factor w/ 110 levels "Antigua Hawksbills",..: 18 18 108 107 24 40 42 88 35 41 ...
 $ winner                          : Factor w/ 105 levels "Afghanistan",..: 17 17 104 103 23 34 14 86 86 44 ...
 $ Home_Away_Winner                : Factor w/ 3 levels "away","Away",..: 2 2 1 2 2 3 3 2 3 2 ...
 $ Bat_F_or_S_Winner               : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 1 1 2 2 ...
 $ Winning_toss_wins_or_Loss_the_match: int  0 1 0 1 1 1 0 0 1 1 ...
 $ win_by_runs                     : int  NA 7 NA 21 20 16 NA NA 23 50 ...
 $ win_by_wickets                  : int  7 NA 1 NA NA NA 7 5 NA NA ...
 $ balls_remaining                 : int  29 NA 0 NA NA NA 7 0 NA NA ...
 $ innings1                        : Factor w/ 111 levels "Afghanistan",..: 98 20 17 109 26 41 50 93 92 49 ...
 $ innings1_runs                   : int  47 131 133 173 140 228 137 139 167 205 ...
 $ innings1_wickets                : int  10 6 8 5 6 7 10 6 6 2 ...
 $ innings1_overs_batted           : num  14.3 20 20 20 16 20 20 17 15 20 ...
 $ innings1_overs                  : num  16 20 20 20 16 20 20 17 15 20 ...
 $ innings2                        : Factor w/ 108 levels "Afghanistan",..: 23 24 106 65 105 49 19 88 43 19 ...
 $ innings2_runs                   : int  48 108 108 80 120 190 105 116 144 106 ...
 $ innings2_wickets                : int  3 9 9 8 10 9 3 5 8 7 ...
 $ Second_Wickets_remaining        : int  7 1 1 2 0 1 7 5 2 3 ...
 $ innings2_overs_batted           : num  11.1 17 15 11 14.1 18 12.5 13 15 14.3 ...
 $ Second_Over_remaining           : num  8.9 3 5 9 5.9 2 7.5 7 5 5.7 ...
 $ innings2_overs                  : num  16 17 15 11 16 18 14 13 15 14.3 ...
 $ Net_run_rate                    : num  3 6.35 7.2 7.27 7.5 ...
 $ D_L_method                      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ target                          : int  NA 116 108 102 NA 207 103 114 NA 157 ...
```

Figure 3: Data in R Studio

Such that removing the unwanted columns which are not useful for our research,removing all the rows which are not important by filtering the data.Cleaned the column which had nwanted formule built in. We have removed all the match which was not played under duck worth lewis method.

3.3 Data Processing

Data conversion is also an important part of data processing where data has been converted to desired format. Like for example we have added a new column in our data set that is net run rate. Basically run rate is a term which is calculated to know the performance of the team.So this column called net run rate can be calculated by formulae run rate=total run scored/total over faced. Like wise we have converted coded value attribue such as
Winning team =1
Losing team=0.

Like wise we have decoded for column Winning toss =1
Losing toss=0

Home team=1
Foreign team =0

We cleaned the data which where repetitive and where not so useful for our research. Sorting and ordering the data has been done for example it has been done with date from Converting the yaml file to .CSV files.

3.4 Machine learning model selection

Now the core part of analysis starts. We are going to use machine and artificial neural network algorithm to predict the out come of the match. The main motive of classification is to classify the unknown input data to the specific output. Machine learning has the most powerful technique to learn the data as we know that there are 2 kinds of learning supervised and unsupervised learning. In our research we are going to use supervised learning algorithms they are as follows *Support vector machine *Decision tree *Random forest *Logistic regression *Multinomial logistic regression *Neural networks

We have chosen these machine learning algorithm because of the following reason
It set a definite distinction between classifiers.
It is very specific about the definition of the classifiers.
We can create a perfect decision boundaries in supervised learning algorithms
We can specifically specify the number of classes we want in our research
After training the machine we dont want the training data again in the memory for any future classification.
It runs in the logarithmic run time.
It can give the accountability of machine learning model using statistical test.
It can handle missing data.
It is very fast at testing times.
Machine learning algorithm clearly classifies the data.

3.5 Class Imbalance problem

Class imbalance problem is a common problem in the machine learning . The class imbalance arises when the class positive data is less then the class negative data there is high probability of getting the class imbalance problem according to Longadge and Dongre (2013).This problem arises in many disciplines they are fraud detection,facial recognition,oil spillage detection,medical diagnosis,anamoly detection, etc.

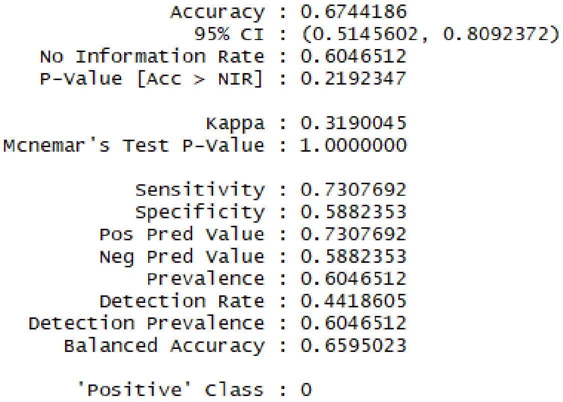
The problem occurs when there is unequal amount of the class present. So machine learning algorithm works well when the classes are roughly equally divided.

When we calculate the specificity and sensitivity of the model we can say that sensitivity percentage value was less then that of the specificity . This indicate that there is a class imbalance in our data sets. To over come this problem or to improve the situation we have used the method of over sampling . We used the Library in the R called ROSE(Randomly over sampling examples). We check the under sampling , Oversampling and both in the function called ovun.sample. We got to know that when we do the undersampling we are able to rightly predict the resource and predicted score more accurately as we can see that our sensitivity increased by undersampling the dataset. When we tried undersampling and oversampling both together we can see that the percentage of accuracy goes slightly down. We also used Synthetic data to overcome the class imbalance for our data set but still the result was not as expected. We have used the resampling method. Sampling is a method to reconstruct our dataset, Including training and validation sets.We got the result good but not as not perfect as of oversampling and undersampling.

4 Implementation

4.1 Random Forest

The first algorithm which we have used is random forest algorithm. This is a supervised learning algorithm is an ensemble learning method.random forest algorithm is built with multitude of decision tree to become a forest.



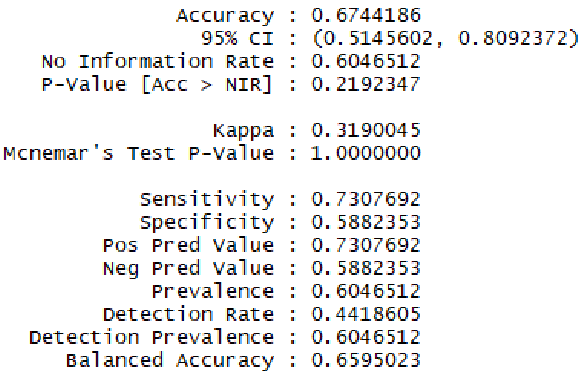
```
               Accuracy : 0.6744186
                 95% CI : (0.5145602, 0.8092372)
    No Information Rate : 0.6046512
    P-Value [Acc > NIR] : 0.2192347

                  Kappa : 0.3190045
 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 0.7307692
            Specificity : 0.5882353
         Pos Pred Value : 0.7307692
         Neg Pred Value : 0.5882353
             Prevalence : 0.6046512
         Detection Rate : 0.4418605
   Detection Prevalence : 0.6046512
      Balanced Accuracy : 0.6595023

       'Positive' Class : 0
```

Figure 4: RF with duck worth lewis method

To perform random forest algorithm we have used the R studio. We installed a library called (randomforest) which is available in the r studio. We do not need to install the package from github.To create the model we splited the data set into 80 percentage training and 20 percentage testing.The main code which we used for our model creation is

rfdUCK¡-randomForest(BatForSWinner SecondWicketsremaining+$SecondOverremaining, data =$
$traindata, importance = TRUE, ntree = 500, do.trace = 100)$
rfold¡-randomForest(BatForSWinner SecondWicketsremaining+$SecondOverremaining+$
$Netrunrate, data = traindata)$
$rfcombination < -randomForest(BatForSWinner$ SecondWicketsremaining+SecondOverremaining
$WinningtosswinsorLossthematch, data = traindata)$

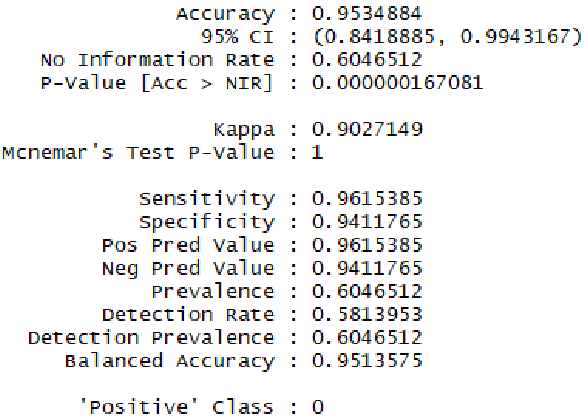
$Fromtheoutputoftherandomforesttreewecanseethatthereisaaccuracyof 0.67 thatmeansof 67 percenta$



```
               Accuracy : 0.6744186
                 95% CI : (0.5145602, 0.8092372)
    No Information Rate : 0.6046512
    P-Value [Acc > NIR] : 0.2192347

                  Kappa : 0.3190045
 Mcnemar's Test P-Value : 1.0000000

            Sensitivity : 0.7307692
            Specificity : 0.5882353
         Pos Pred Value : 0.7307692
         Neg Pred Value : 0.5882353
             Prevalence : 0.6046512
         Detection Rate : 0.4418605
   Detection Prevalence : 0.6046512
      Balanced Accuracy : 0.6595023
```

Figure 5: Rf with Run rate

The above figure shows the calculated result of random forest model for duckworth lewis method with the old method. That means Run rate has been included in this model



```
               Accuracy : 0.9534884
                 95% CI : (0.8418885, 0.9943167)
    No Information Rate : 0.6046512
    P-Value [Acc > NIR] : 0.000000167081

                  Kappa : 0.9027149
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9615385
            Specificity : 0.9411765
         Pos Pred Value : 0.9615385
         Neg Pred Value : 0.9411765
             Prevalence : 0.6046512
         Detection Rate : 0.5813953
   Detection Prevalence : 0.6046512
      Balanced Accuracy : 0.9513575

       'Positive' Class : 0
```

Figure 6: Rf with improved vesion

The above model is for duck worth lewis method with old and improved version. This model shows the highest accuracy of 95 percentage.

4.2 Decision tree

Decision tree is a decision tool system. which when created it looks like a tree structure with its root in the top side.This model more or less is a graph like structure with inter connected nodes. To create the model we split the data into training and testing into 80 percentage and 20 percentage respectively.

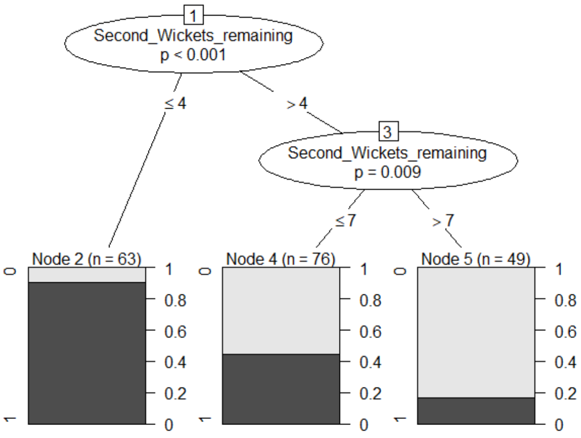


Figure 7: Decision Tree

We used following code to create the decision tree model that is for duck worth lewis, duck worth lewis old model and duck worth lewis improved version.

Decisionmodel¡-ctree(BatForSWinner SecondWicketsremaining+$SecondOverremaining, data = tdata, controls = ctreecontrol(mincriterion = 0.9, minsplit = 50))$

Decisionmodelold¡-ctree(BatForSWinner SecondWicketsremaining+$SecondOverremaining + Netrunrate, data = tdata, controls = ctreecontrol(mincriterion = 0.9, minsplit = 50))$

Decisionmodelcombination¡-ctree(BatForSWinner SecondWicketsremaining+$SecondOverremainin$
$Netrunrate + WinningtosswinsorLossthematch, data = tdata, controls = ctreecontrol(mincriterion$
$0.9, minsplit = 50))$

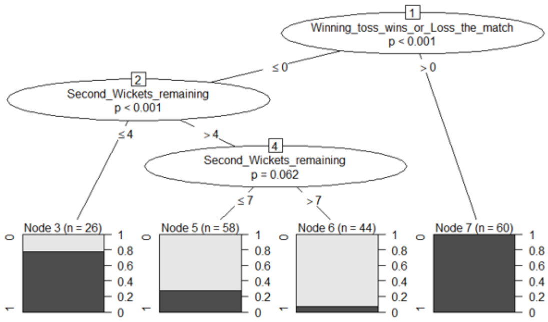


Figure 8: Decision tree With improved version

To create the above model we have used library package called party in the R studio. To create the model we used the controls such as ctreecontrol with the mincriterion value of 0.9, and minsplit of 50 percentage.The same value of mincriterion and minsplit has been given to all the model.

4.3 Support vector machine

support vector machine is a supervised learning model.An SVM model is a probabilistic binary linear classifier.To build a model we have used a library called e1071 of R language.

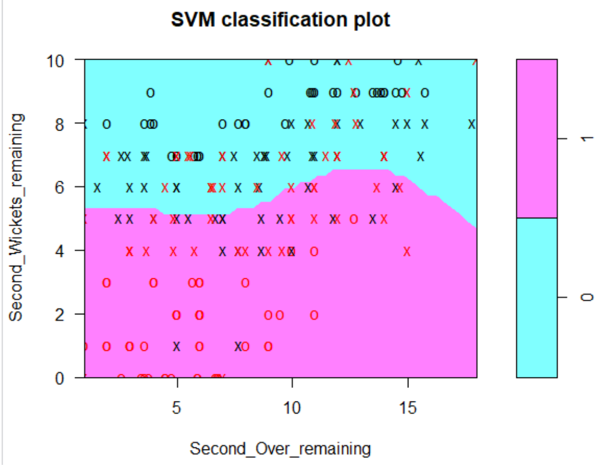


Figure 9: support Vector machine

To create the model we have used the following code svmmodelduck¡-svm(BatForSWinner SecondWicketsremaining+$SecondOverremaining, data = traindata$) svmmodelold¡-svm(BatForSWinner $SecondWicketsremaining+SecondOverremaining+ Netrunrate, data = traindata$) svmmodel¡-svm(BatForSWinner SecondWicketsremaining+$SecondOve$ $Netrunrate + WinningtosswinsorLossthematch, data = traindata$)

The data has been split into 80 and 20 percentage of train and test data.

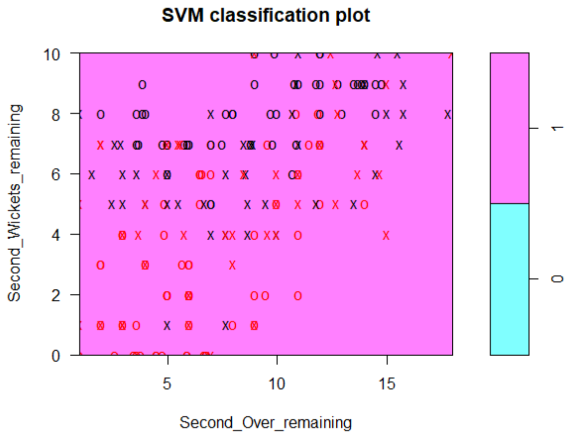


Figure 10: Support vector machine with the Improved version

To Plot the model of support vector machine we have used the plot function with slice of list as a function variable. We have divided the split into net run rate=3 and win toss and win loss match =4.

4.4 logistic regression

Logistic regression or we can say as logit regression.Here to create a model we need a dependent and independent variable.To create this model we have used a glm model. The funtion which is used to create the model is

Logisticmodel¡-glm(BatForSWinner SecondWicketsremaining+$SecondOverremaining, data = tdata, family = binomial)$
Logisticmodelold¡-glm(BatForSWinner SecondWicketsremaining+$SecondOverremaining + Netrunrate, data = tdata, family = binomial)$
Logisticmodelcombination¡-glm(BatForSWinner $SecondWicketsremaining+SecondOverremainin$
$Netrunrate + WinningtosswinsorLossthematch, data = tdata, family = binomial)$

To create this model family variable to which we have given a value of binomial. We have used 80 and 20 percentage of ratio for training and testing the data.

To find out the mis-classification error in this model we used the 43 observations.These observations are found out during inspection of the date. This 43 observation are the main data which help us to build the Logistic model.

4.5 Multinomial logistic regression

Multinomial logistic regression is classification method . This algorithm generalizes the logistic regression.This algorithm can produce more then two different outcomes.To create this model first we changed our dependent variable into the factor variable. This is first condition we need to follow if we want to build a Multinomial logistic model. For data set we have divided the data into 80 and 20 percentage for training and testing the data.Before using the training and testing the data we have re-leveled the dependent variable.Re-level has done in order to form a matrix with ordered value.The model which we have created using following code.

mymodeld¡-multinom(out SecondWicketsremaining+SecondOverremaining,data=CricketCsv)
mymodelold¡-multinom(out SecondWicketsremaining+$SecondOverremaining+Netrunrate, data =$
$CricketCsv)$
mymodelcombination¡-multinom(out SecondWicketsremaining+SecondOverremaining+Netrunrate-
$CricketCsv)$

To create the above model we have used the library called nnet.To predict the model we have used the Type="Prob"

4.6 Neural network

Artificial neural network or ANN is also called as a connectionist system. This algorithm is inspired by biological neural system.To create Artificial neural network model we need to change a data in a certain way that the values are between 0 and 1 . To achieve this we use normalization . We used the Min -Max normalization to all the independent variable which are needed for our model creation . The independent variable which we have chosen to create this model are Netrunrate,SecondWicketsremaining,SecondOverremaining,WinningtosswinsorI

To create the model we have dropped the column which are not necessary for our model creation. we have used the library called neural net for our model. Model is created using following code.

neuralmodelold¡-neuralnet(BatForSWinner SecondWicketsremaining+$SecondOverremaining+Netrunrate, data = Traindata, hidden = 2, err.fct = ”ce”, linear.output = FALSE)$

neuralmodelold¡-neuralnet(BatForSWinner SecondWicketsremaining+$SecondOverremaining+Netrunrate+WinningtosswinsorLossthematch, data = Traindata, hidden = 2, err.fct = ”ce”, linear.output = FALSE)$

neuralmodel¡-neuralnet(BatForSWinner SecondWicketsremaining+$SecondOverremaining, data = Traindata, hidden = 2, err.fct = ”ce”, linear.output = FALSE)$

Here we have splited the data into 80 and 20 percentage for our training and testing data respectively.

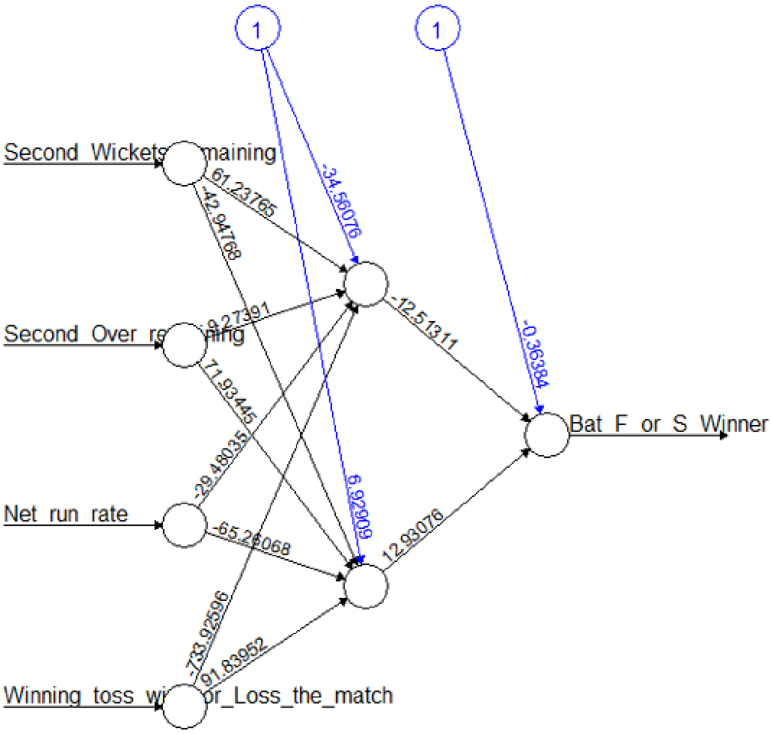


Figure 11: Neural network with two hidden layer

Above figure is the neural network with 2 hidden layer.Two create this model we can see that from left side we have independent variable this are second wicket,second over,net run rate and Winning toss loss the match. These are variables used with dependent variable as the output bat f or s winner with the value of 0.383. An our dependent variable with 61.23,71.93,65.26 and 91.93

5 Evaluation

In this section we are going to evaluate the result of three aspect of our research and going to compare.These are Duck worth lewis method, Duck worth lewis method and Run rate and duck worth lewis method with the improved version.

5.1 Performance of Duck worth lewis method / Case Study 1

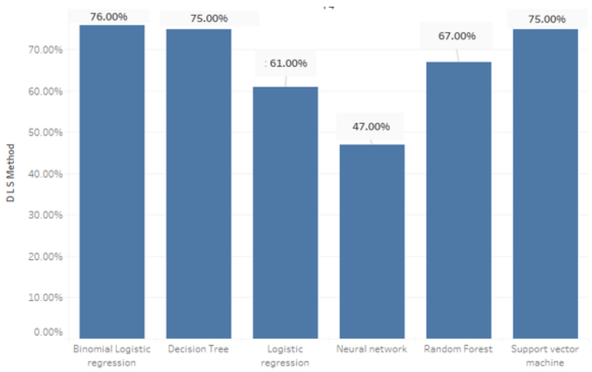


Figure 12: performance of duckworth Lewis

We need to note that for all our results in the percentage has been rounded off.For this result we have added two factor to see the performance of our model these are wicket remaining and overs remaining. As we can see from the result that Binomial logistic regression has the highest accuracy of 76.00 percentage with decision tree and support vector machine to be second . We can note here that neural network performs very low accuracy of 47.00 percentage.

5.2 Duck worth lewis method with the old method(Run rate)/ Case Study 2

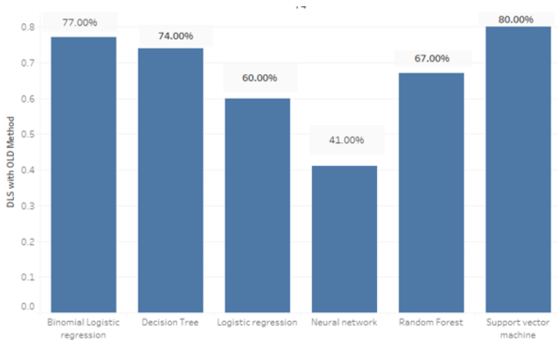


Figure 13: Performance of D/L Method with Run rate

.

To see the performance of the duck worth lewis method with old method added in them we have added used three variables they are wickets remaining, over remaining and run rate of the match . As we can see from the result that support vector machine shows the best result 80 percentage with Binomial logistic regression,decision tree,Random forest, and logistic regression stands second, third ,fourth and fifth.Here in this case we can also see that the performance of the neural network is worst for this case under two hidden neurals In ANN.

5.3 Duck worth lewis method with old and improved version/ Case Study 3

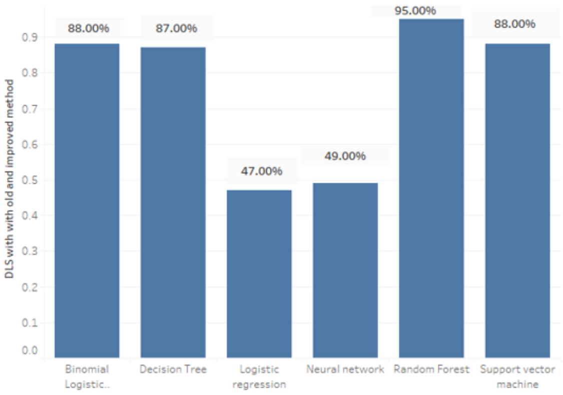
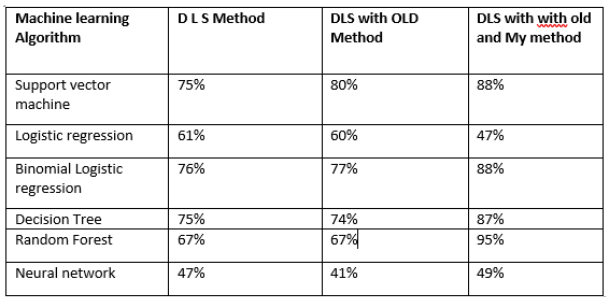


Figure 14: Performance of improved Duckworht lewis method

This is the result of the improved duckworth lewis method we have added 4 variables in this model to check its performance.Here we can see that the performance of Support vector machine,Binomial logistic regression,Decision tree,Random forest and Neural network is very high when we compare with above two models. Here we can note that the performance of logistic regression is very low as compare to other two models.In this model we can see that Random forest method perform best among all the algorithm and Logistic regression performs worst.

5.4 Conclusion / Case Study N



Machine learning Algorithm	D L S Method	DLS with OLD Method	DLS with with old and My method
Support vector machine	75%	80%	88%
Logistic regression	61%	60%	47%
Binomial Logistic regression	76%	77%	88%
Decision Tree	75%	74%	87%
Random Forest	67%	67%	95%
Neural network	47%	41%	49%

Figure 15: Overall Analysis

Here we can conclude from the result that the accuracy of our model as compare with the other models are best . We can see that Support vector machine,Binomial Logistic regression,Decision tree,Random forest, and neural network performs best with the improved version of duck worth lewis method and with the logistic regression we have the worst performance index. So we conclude that as the related variable such as run rate and Win and loss variables are added into our model the accuracy of all model increases except for the logistic regression.

...

6 Conclusion and Future Work

The main purpose of this research is to improve the duck worth lewis method using machine learning algorithm and models. We have also compared the Duck worth lewis method, Duck worth lewis method with run rate and improved version of the duck worth lewis method. Here we can conclude that we have successfully improved the method using artificial neural network,decision tree, Random forest, binomial logistic regression. But we fail to achieve the performance using the logistic regression. From the result above we can say that Duck worth lewis method performance is much lesser then the other two proposed methods.The improved version of Duck worth lewis method is much efficient and there accuracy rate is much higher then the other two proposed model. The main limitation of our research is that we are not able to achieve the better result using the logistic regression.For future work to improve the model we can use the other factors such as Winning toss, Pitch and home team as new variable to add into the existing model to test the improved version of duck worth lewis method.We can also add new factor such as strength of the team. But researcher could face a problem with defining the strength of the team because we cannot represent the strength with mathematical formula. We can also choose some deep learning algorithm to compare the data mining neural network and deep learning algorithm.

References

Bandulasiri, A. (2008). Predicting the winner in one day international cricket, *Journal of Mathematical Sciences & Mathematics Education* **3**(1): 6–17.

Beaudoin, D. and Swartz, T. (n.d.). One-day cricket.

Hucaljuk, J. and Rakipović, A. (2011). Predicting football scores using machine learning techniques, *MIPRO, 2011 Proceedings of the 34th International Convention*, IEEE, pp. 1623–1627.

Joseph, A., Fenton, N. E. and Neil, M. (2006). Predicting football results using bayesian nets and other machine learning techniques, *Knowledge-Based Systems* **19**(7): 544–553.

Kaluarachchi, A. and Aparna, S. V. (2010). Cricai: A classification based tool to predict the outcome in odi cricket, *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on*, IEEE, pp. 250–255.

Longadge, R. and Dongre, S. (2013). Class imbalance problem in data mining review, *arXiv preprint arXiv:1305.1707* .

Martins, R. G., Martins, A. S., Neves, L. A., Lima, L. V., Flores, E. L. and do Nascimento, M. Z. (2017). Exploring polynomial classifier to predict match results in football championships, *Expert Systems with Applications* **83**: 79–93.

McHale, I. G. and Asif, M. (2013). A modified duckworth–lewis method for adjusting targets in interrupted limited overs cricket, *European Journal of Operational Research* **225**(2): 353–362.

Me, E., Unold, O. et al. (2011). Machine learning approach to model sport training, *Computers in human behavior* **27**(5): 1499–1506.

Pathak, N. and Wadhwa, H. (2016). Applications of modern classification techniques to predict the outcome of odi cricket, *Procedia Computer Science* **87**: 55–60.

Phanse, V. and Deorah, S. (2011). Evaluation and extension to the duckworth lewis method: A dual application of data mining techniques, *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, IEEE, pp. 763–770.

Ramanayaka, A., Semansinghe, W., Ohani, P. and Wehigaldeniya, W. (2016). Application of econometrics in sport: A probability estimation of getting outin odi cricket, Sri Lanka Forum of University Economists (SLFUE), Department of Economics, Faculty of Social Sciences, University of Kelaniya.

Sankaranarayanan, V. V., Sattar, J. and Lakshmanan, L. V. (2014). Auto-play: A data mining approach to odi cricket simulation and prediction, *Proceedings of the 2014 SIAM International Conference on Data Mining*, SIAM, pp. 1064–1072.

Singh, T., Singla, V. and Bhatia, P. (2015). Score and winning prediction in cricket through data mining, *Soft Computing Techniques and Implementations (ICSCTI), 2015 International Conference on*, IEEE, pp. 60–66.

Xu, B. (2012). Prediction of sports performance based on genetic algorithm and artificial neural network, *International Journal of Digital Content Technology and its Applications* **6**(22): 141.