Predicting a T20 cricket match result while the match is in progress

Authors Name

Fahad Munir (11201014)
Md. Kamrul Hasan (11201032)
Sakib Ahmed (11201009)
Sultan Md. Quraish (11201017)

Supervisor

Rubel Biswas

Co-Supervisor

Moin Mostakim

A thesis presented for the degree of
Bachelor in Computer Science

Inspiring Excellence

Department of Computer Science and Engineering
BRAC University, Bangladesh
23/8/2015

Acknowledgement

Firstly, all praise to the Great Allah for Whom our thesis have been completed without any major interruption.

Secondly, to our advisor Mr. Moin Mostakim sir for his kind support and advice in our work. He helped us whenever we needed help.

Thirdly, Jon Van Haaren and the whole judging panel of Machine Learning in Sports Analytics Conference 2015. Though our paper not accepted there, all the reviews they gave helped us a lot in our later works.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

1 Abstract

Data Mining and Machine learning in Sports Analytics, is a brand new research field in Computer Science with a lot of challenge. In this research the goal is to design a result prediction system for a T20 cricket match while the match is in progress. Different machine learning and statistical approach were taken to find out the best possible outcome. A very popular data mining algorithm, decision tree were used in this research along with Multiple Linear Regression in order to make a comparison of the results found. These two model are very much popular in predictive modeling. Forecasting a T20 cricket match is a challenge as the momentum of the game can change drastically at any moment. As no such work has done regarding this format of cricket, we have decided to take the challenge as T20 cricket matches are very much popular now a days. We are using decision tree algorithm to design our forecasting system by depending on the previous data of matches played between the teams. This system will help the teams to take major decision when the match is in progress such as when to send which batsman or which bowler to bowl in the middle overs. It significantly expands the exposure of research in sports analytics as it was previously bound between some other selected sports.

2 Abbreviations

ODI: One day International

T20: Twenty Twenty

IPL: Indian Premier League

CPL: Caribbean premier League

BPL: Bangladesh Premier League

MR: Runs scored by Home team

OR: Runs scored by the opponent team

MRN: Home Team Run Rate

ORN: Opponent Team Run Rate

Batord: Batting Order

MW: Home Team Wicket

OW: Opponent Team Wicket

MAE: Mean Absolute Error

RMSE: Root Mean Squared Error

LBW: Leg Before Wicket

NBA: National Basketball Association

CSK: Chennai Super Kings

Table of Contents

List of Figures

List of Tables

3 Introduction

There have been research done on ODI and Test match cricket but very few on T20 cricket, which is currently more favourite than its older brothers. And that's why we decided to di research on this format of the game. The result of a T20 cricket match depends on lots of in game and pre-game attributes. Pre-game attributes like condition, venue, pitch, team strength etc. and in game attributes like wickets in hand, run rate, total run, strike rate etc. influence a match result predominantly. We gave more emphasis on in game attributes as our prediction will be when match is in progress. Our intentions would be to finding out the attributes which is most affecting the result in different phases of the game. We broke an innings into three phases: Power-play (1-6 overs), Mid-overs(7- 16) and final overs(17-20). Prediction will be active till the last over of mid overs phase. We consider an entire cycle of process of data mining, decision making and preparing a model to predict. Mining the data according to the attributes and different phases we have divided important to construct meaningful statistics. Modeling a problem for prediction requires several intelligent assumptions and molding the problem with collected data-sets. As we already mentioned cricket is a game of uncertainty and T20 format is the most unpredictable format rather than the other two format because it is the shortest format of the game and one over can change the result of a game.

In this research we tried to design a prediction model which can go with this unpredictability and try to make a result prediction.

3.1 Aims and Objectives

The aim is to prepare a model which will predict the result of a T20 cricket game while the match is in progress. Our main objective is to combine pre-game data and in-game data in order to design a good predictive model. Understanding the different attributes is also needed in order to get more accuracy in result.

3.2 History of T20 Cricket

From the sixteenth century to first official Test match in 1877 to first ODI World Cup 1975 with 60 overs to 50 overs world in 1987, cricket world has changed a lot. Now, we have a new shortest format in cricket - Twenty20 or T20. T20 cricket is fun, entertaining and more thrilling than other two formats. It has brought glamour and instant popularity to the fans and helped marketing Cricket to the rest of the world. England Cricket Board (ECB) was looking for a cricket competition to fill the void after the conclusion of Benson and Hedges Cup in 2002. ECB was looking for something new to attract more sponsors and viewers. Marketing Manager of ECB, Stuart Robertson was the first to came up with the idea of playing cricket match with each team getting only 20 overs to play. Thus came the name Twenty - 20. First official T20 matches were played in English counties. Though first official T20 international match took place between Australia and New Zealand. In 2007 we witnessed ICC World Twenty20, which generated immense support for this newest form of cricket. Introduction of T20 cricket gave birth to franchise league in many countries. Among those Indian Premier League(IPL) is the most watched and expensive cricket league. Big Bash, Caribbean Premier League(CPL) and Bangladesh Premier League(BPL) other popular franchise leagues. T20 cricket showed great innovation in batting style, improved fielding. Bowlers were also trying their hardest to make them useful in a format which was made to give preference to the batsmen. With more viewers and sponsors, T20 cricket brought more money to the Boards and players. But this format also attracted more illegal activities as matching fixing, betting, miss conduct of players. Very recently in July 2015 BCCI banned Chennai Super Kings and Rajasthan Royals for match fixing. Ironically this two teams are two of the most successful team in IPL with a large number of fan base. The West Indies regional teams completed in tournament named Stanford 20/20 which was funded by a convicted fraudster Allen Stanford.

3.3 Game Method

After Football, Cricket is the second most popular sports with a fan base of around 2.5 billion (according to Top End Sports) and mostly popular in South Asia, Australia, The Caribbeans and UK. In international level Cricket is played in three formats- Test, ODI and T20I cricket. This game is played on a 22 yards clay pitch with 2 sets of stamps, each set with 3 stamps and each set having two bells on top of them. Two batsmen come to pitch with two wooden bats and bowler bowls with a cricket ball which outer part is made of lather. Test Cricket is played Red ball which is slightly heavier than the White bowl played in the limited overs. There is no fixed size of the outfield, but usually its diameter usually varies between 137 meters and 150 meters. In limited over cricket there is a circle of 30 yards around the pitch which work as a field restriction for players. Test cricket is played for 5 days with each team having at most 2 innings. ODI played for 50 overs per innings and T20 played in 20 overs. Each team play with 11 players. A coin toss decides who is going to bat or ball first. In limited over cricket team batting first scores as many run possible before the overs are finished or they all get out. If team batting next score more runs they wins and failure to score required runs in allocated overs or getting all out result in loss for team batting second. Some basic idea how the game is played:

– **Field Restriction**: According to the latest rule change in 50 overs cricket, there is only one Powerplay from over 1-10 with only two fielders outside of the 30 yards circle. between 11 to 40 overs four fielder are allowed and five allowed outside the 30 yards circle in the final 10 overs. Like the ODI format T20 also have only 1 powerplay form over 1 to 6 with 2 fielders outside the circle.

- **Scoring Runs**: The striking batsman must hit the ball with his bat and must change his position with his partner to score 1 run. Number of runs scored depend on the number of time the batsmen change position. If the batsman hit ball and its goes outside the boundary 4 runs are added and 6 runs are added when the ball fly over the boundary. Batting team gets extra runs form No ball, Leg bye, Bye, Wide, Overthrows and Penalty runs when the ball hits keepers helmet or cap lying on the field.
- **Out Types:**Batsmen usually get out by being bowled, caught, leg before wicket(LBW), stumped and run out. There are some rare occasion where batsmen get out by hit wicket, intentionally hitting the ball twice, handled the ball, obstructing the field and timed out.
- **Tie match result**:If the match is tie, such as both the team scored same runs then there is a rule.It's called super over. Super over played for only one over for each team. Each team can play with two wickets when they are batting and one single bowler when they are bowling. Batting first team set a target and second team chase it.

In Test cricket there is no restriction on how many overs a bowler can bowl. But in limited over crickets number of overs bowled by a single bowler is fixed. in ODI's each bowler can bowl up to 10 overs in a match and in T20 cricket bowlers are allowed to bowl only 4 overs each.

4 Related Work

Better predictive modeling depends on better understanding of the data and attributes selection. We have to choose between some data mining algorithm. We have chosen data mining as it is very much flexible in predictive modeling. Prediction when the game is in progress is a tough ask and it need finding the best attributes that influence the match outcome. Some research was done previously on predictive modeling in sports like Basketball, Baseball along with Test and One Day International cricket.

In basketball, Bhandari et al.[2] developed a knowledge discovery system and data mining framework for National Basketball Association (NBA). It was aimed to discover several interesting patterns in basketball games. This and related system have been used by several basketball teams over the past decades. Such solutions designed for offline usage and no in game effects were taken care of. There has been some recent works (20) about in-game decision making to find how much time remaining in the game without making any prior prediction model.

There were several works done in cricket. Bailey and Clarke [4] and Sankaranarayanan et al.[1] used machine learning approach to predict the result of a one day match depending on the previous data and in game data.

Akhtar and Scarf [7] used multinomial logistic regression in their work on predicting a outcome of a test matches played between two teams.

Choudhury et al.[8] used Artificial Neural Network to predict result of a multi team one day cricket tournament depending on the past 10 years data. They used training set in order to model the data in neural network. Again there was no in play effects were taken care of.

For baseball, Ganeshapillai and Guttag [9] developed a prediction model that decides when to change the starting pitcher as the game progresses. It is very much similar to our work-flow, where they used the combination of previous data and in game data to predict a pitchers performance.

Tulabandhula and Rudin [6] were designed a real time prediction and decision system for professional car racing. Model makes the decision of when is the best time for tire change and how many of them. These works supplied a huge encouragement and informative ideas in our research.

5 Tools and Softwares

5.1 Weka

Weka is an intelligent data mining tool with the support of analyzing different data mining algorithm. As for the decision tree algorithm, it takes the dataset in .arff format. After analysing the data it gives us the decisions in a tree format. In weka we can also import data from a database in order to be more flexible handling of data. Beside the decision tree, it also show the accuracy, correctly and incorrectly handled data, mean absolute error, root mean squared error, relative absolute error, kappa statistics.

Mean absolute error(MAE) The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. The equation is given in the library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

Root mean squared error(RMSE) The RMSE is a quadratic scoring rule which measures the average magnitude of the error. The equation for the RMSE is given in both of the references. Expressing the formula in words, the difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude.

6 Data Description

Collection of a large dataset is the main perspective of a data mining and machine learning research and collecting data with proper understanding from reliable source is most important. In our research, we needed a strong and reliable data source which we found in Statsguru. Statsguru is ESPN Cricinfo's cricket statistics maintenance database, where all the data relating to all the matches of cricket are saved. In this database system, all the match's data are stored with live ball by ball commentary. There are lots of different kinds of formats of data and statistics however we have our own format of data collection in order to design our predictive system.

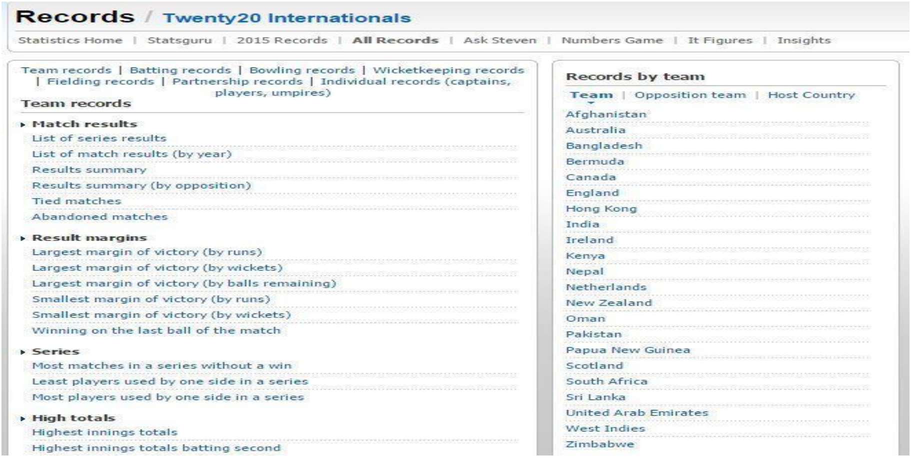
In our data-sets we have chosen a team as a home team. Depending on that we have divided the match attributes. The attributes are: Venue, runs of home team in a segment (MR), runs of opposition team in a segment (OR), run rate of home team in a segment (MRN), run rate of opposition team in a segment (ORN), teams batting order (which team batted first, which team second), number of wickets fallen for home team (MW), number of wickets fallen for opposition team (OW) and finally the result of the match. Statsguru's system is well organized and quite easy to understand. Information were perfectly grouped in different sections which shows different kinds of statistics about team and individual records.

On the right side of the website they have a tab called 'Records', where data are divided into different divisions and subdivisions such as Test Matches, One Day Internationals, T20 Internationals as well as several domestic competitions like Indian Premier League (IPL), Caribbean Premier League (CPL), Bangladesh Premier League (BPL), England's Natwest T20 blast, Australia's Big Bash T20 league with some other domestic First Class (4-day matches such as Ranji trophy) and List A (One day 50 over matches) tournaments.

From all of these information, we have collected the T20 International match data along with Indian Premier League data.

Till August 16, 2015 number of international T20 cricket matches is 452. Number of IPL matches played is 524. Also a lot of matches were played in other domestic and franchise leagues. As our match prediction model can perform on two playing teams, we need data from the matches played previously by this two teams. From the international T20 part, we added data from Australia vs other Test playing countries. Australia has played 77 T20I till now. Besides, we also took data from IPL. We collected match details of Chennai Super Kings (CSK) as the Home team and took the matches between CSK vs Mumbai Indians, CSK vs Rajasthan Royals, CSK vs Kolkata Knight Riders, CSK vs Kings XI Punjab CSK vs Royal Challengers Bangalore and collected CSK vs Delhi Daredevils to test our prediction. Data collection started from the year the IPL season 1 to season 8. We collected around 200 T20 matches data and did our analysis on them. We started collecting our data from the International T20 matches. Firstly, match data relating to Australia were done. T20 International part have different subdivisions named as Match results, Results margin, Series and so on. Besides, Cricinfo website has a very powerful statistical analysis of the previous games. One question may arise that is there data is reliable for research purpose. We have been working on this website almost for a year, but till now we have not found any mistake or misleading information here.

As we were looking for match by match results of a team, we started with Australia vs Pakistan matches. We took Australia as the home team. In fig: 1 if we look at the right side there is record by teams.



Records / Twenty20 Internationals

Statistics Home | Statsguru | 2015 Records | All Records | Ask Steven | Numbers Game | It Figures | Insights

Team records | Batting records | Bowling records | Wicketkeeping records | Fielding records | Partnership records | Individual records (captains, players, umpires)

Team records

- Match results
  - List of series results
  - List of match results (by year)
  - Results summary
  - Results summary (by opposition)
  - Tied matches
  - Abandoned matches
- Result margins
  - Largest margin of victory (by runs)
  - Largest margin of victory (by wickets)
  - Largest margin of victory (by balls remaining)
  - Smallest margin of victory (by runs)
  - Smallest margin of victory (by wickets)
  - Winning on the last ball of the match
- Series
  - Most matches in a series without a win
  - Least players used by one side in a series
  - Most players used by one side in a series
- High totals
  - Highest innings totals
  - Highest innings totals batting second

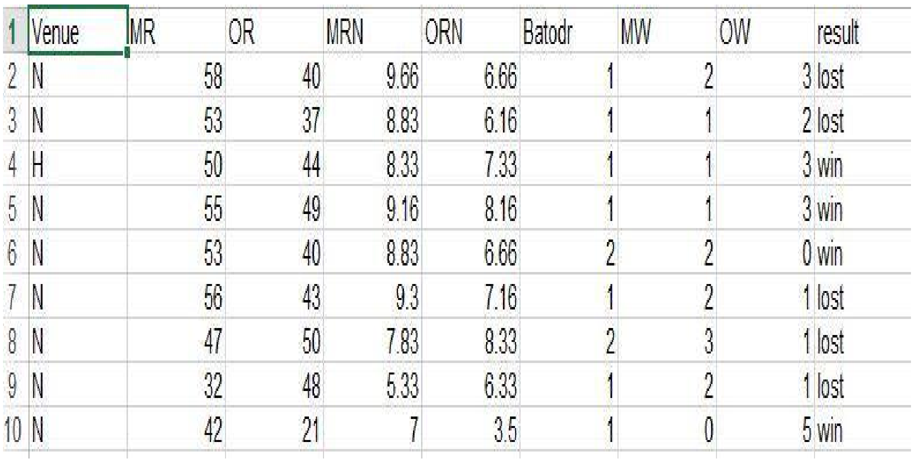
Records by team

Team | Opposition team | Host Country

Afghanistan
Australia
Bangladesh
Bermuda
Canada
England
Hong Kong
India
Ireland
Kenya
Nepal
Netherlands
New Zealand
Oman
Pakistan
Papua New Guinea
Scotland
South Africa
Sri Lanka
United Arab Emirates
West Indies
Zimbabwe

Fig. 1. Data Collection

We choose Australia as the home team, then from the list of match results we have year by year data of Australia. We are going to take a match between Australia and Pakistan that played in 10 September 2012, number 10 row as an example to elaborate our work. We opened the match with the Pakistan and went to the over comparison. Attributes that mentioned above, we started filling those attributes with proper data accordingly. Firstly, we collected data from 0-6 overs. In the first attribute named as Venue is ‘N’ (N represents neutral Venue) because the match played in Dubai. It is not home ground for any team. If Australia played the match in their country, then we would mention the venue as ‘H’ (H represents Home venue) and if the match was played in Pakistan then the Venue would have been ‘A’ (A represents Away Venue). Secondly, Australia batted first so the ‘Batord’ attribute is 1, on the other hand it would be 2 if they batted second. The result part represents the end result of the match, in this circumstances Australia won the match.

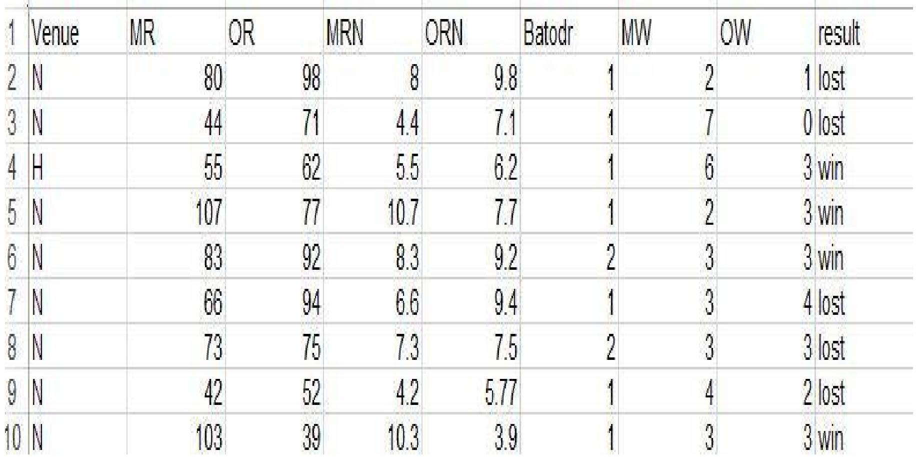


1	Venue	MR	OR	MRN	ORN	Batodr	MW	OW	result
2	N	58	40	9.66	6.66	1	2	3	lost
3	N	53	37	8.83	6.16	1	1	2	lost
4	H	50	44	8.33	7.33	1	1	3	win
5	N	55	49	9.16	8.16	1	1	3	win
6	N	53	40	8.83	6.66	2	2	0	win
7	N	56	43	9.3	7.16	1	2	1	lost
8	N	47	50	7.83	8.33	2	3	1	lost
9	N	32	48	5.33	6.33	1	2	1	lost
10	N	42	21	7	3.5	1	0	5	win

Fig. 2. Dataset of first segment

Now the part of MR (Runs scored by Home team) in this 0-6 overs Australia took 42 runs and the OR (Runs scored by the opponent team) was 21. Then we divided these runs with 6 because they completed playing 6 overs of the match and so the MRN (Home Team Run Rate) we put 7 as the run rate and 3.5 in ORN (Opponent Team Run Rate). Following that there is MW (Home team Wicket) is 0 because Australia did not lose any wicket within these overs but Pakistan lose five wickets.

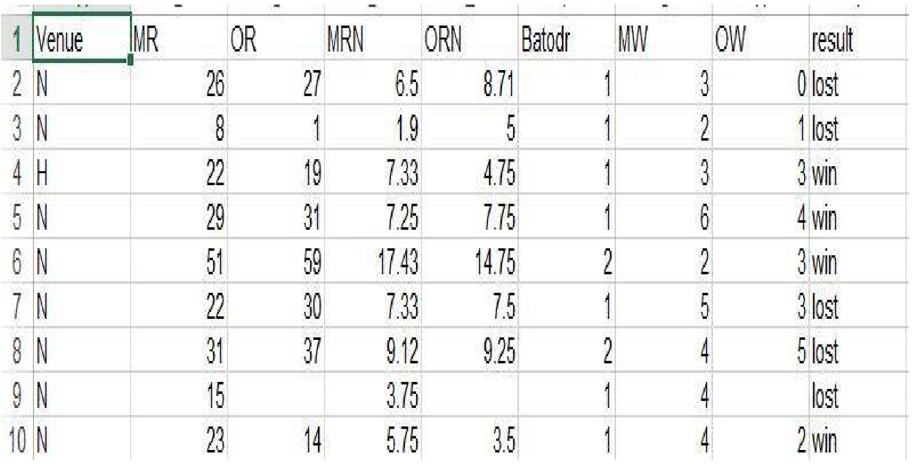
Secondly, in the next segment we took 10 over from 7-16 of the match. In order to calculate the total scored run MR by substituting the current run from the run we have till the sixth over. Same procedure followed for the OR. Both the MRN and ORN calculated by dividing the MR and OR by 10 for the reason that it is for total 10 overs. From example, we can see that Australia took 103 runs from 7-16 over whereas Pakistan only managed to score 39 runs in these 10 overs. The MRN and ORN was 10.3 and 3.9 respectively. The wicket fall in these 10 over period was 3 wickets which is same for the both team.



1	Venue	MR	OR	MRN	ORN	Batodr	MW	OW	result
2	N	80	98	8	9.8	1	2	1	lost
3	N	44	71	4.4	7.1	1	7	0	lost
4	H	55	62	5.5	6.2	1	6	3	win
5	N	107	77	10.7	7.7	1	2	3	win
6	N	83	92	8.3	9.2	2	3	3	win
7	N	66	94	6.6	9.4	1	3	4	lost
8	N	73	75	7.3	7.5	2	3	3	lost
9	N	42	52	4.2	5.77	1	4	2	lost
10	N	103	39	10.3	3.9	1	3	3	win

Fig. 3. Dataset of Second Segment

Lastly, in the final overs (17-20), its get a little trickier as it is the final segment of the game. In the first innings batting team wants team wants to score as many runs as possible to make a huge total to chase for the opposition team. The team batting second gets all out or chase down the required target. This scenario sometimes occurs in middle over too. When this happens we adjust our calculations accordingly. The instance we are using represents that Pakistan was all-out within 19.1 overs, meaning they have not played the whole 4 overs. So while calculating the result we were careful to divide the run with 3.1 overs, as it gave us the proper OMR value.



1	Venue	MR	OR	MRN	ORN	Batodr	MW	OW	result
2	N	26	27	6.5	8.71	1	3	0	lost
3	N	8	1	1.9	5	1	2	1	lost
4	H	22	19	7.33	4.75	1	3	3	win
5	N	29	31	7.25	7.75	1	6	4	win
6	N	51	59	17.43	14.75	2	2	3	win
7	N	22	30	7.33	7.5	1	5	3	lost
8	N	31	37	9.12	9.25	2	4	5	lost
9	N	15		3.75		1	4		lost
10	N	23	14	5.75	3.5	1	4	2	win

Fig. 4. Dataset of Final Segment

7 Prediction Modeling using Decision Tree

Decision tree algorithm is a very popular way to design a predictive modeling.Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Runs ,Wickets and Run-Rate). Leaf node (e.g., Result) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree built on the calculation of Entropy and Information gain.

7.1 Entropy Calculation:

Entropy is a measure of unpredictability and uncertainty of a data-set. Entropy is generally considered to determine how disordered a data-set is. The higher rate of entropy refers to the uncertainty and more information needed in these cases to improve the predictability. One outcome is very much certain when the entropy is zero.

$$Entropy(S) = \sum_{i=1}^{C} P_i log_2 P_i \tag{1}$$

where Pi is the proportion of instances in the dataset that take the i-th value of target attribute, which has C different values. This probability measure give us the idea of how uncertain we are about the data. We use a log_2 measure as this represents how many bit we would need in order to specify what the class is of a random instance.

7.2 Information Gain:

Now we want quantitative way of splitting the data-set by using a particular attribute. We can use a measure called Information Gain, which calculates the reduction in entropy that would result in splitting the data on an attribute, A. Information Gain is actually a procedure to select the particular attribute to be a decision node of a decision tree.

$$Gain(S,A) = Entropy(S) - \sum_{v \epsilon A} \frac{S_v}{S} Entropy(S_v) \tag{2}$$

where v is a value of A, Sv is the subset of instances of S where A takes the value v and S is the number of instances With the help of this node evaluation technique we can proceed recursively through the subset we create until leaf nodes have been reached throughout and all subsets are pure with zero entropy. This is how a decision tree algorithm works.

7.3 Data Training

After collecting the data we converted those data into an attributed relation file format(arff) and then we have used weka for classification. After classification using some algorithm we got some result and later we have analyzed those result.Here is the simple work flow chart given.

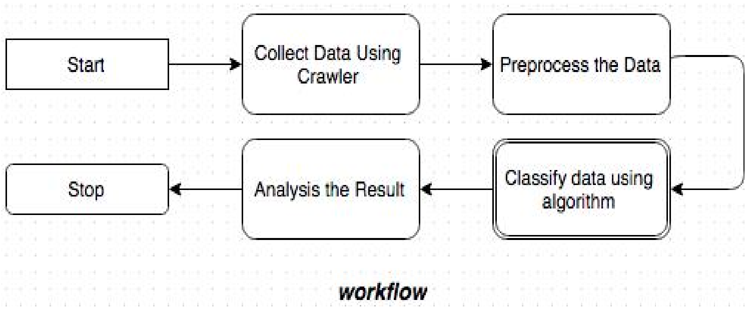


Fig. 5. Workflow

8 Analysis and Result

As we have divided our total work into 3 segments . First of all, we selected our first 6 over data set and we have implemented the algorithm over it. Then we have analyzed the data visually and figure out the distribution of values. The following metrics were used to determine the performance of our model: Time taken to build the model, Kappa statistics, mean absolute Error, Root Mean Squared

Error, Relative Absolute Error, Prediction accuracy.

8.1 Win Prediction

Bat First: In our modeling, we used all the matches played by Chennai Super Kings (CSK) against the other teams in IPL. All the data classified for training set and test set are from that data-set. In this batting first part, we are showing that the winning chance of CSK when batting first.Sample data-set were matches played between Chennai Super Kings and Mumbai Indians in IPL history. Here is the decision tree:

Evaluation of Training Set: Below figure showing the Accuracy of the training set for the batting first. We used 70 % of the data to create a training set which gives us the accuracy of 78 %. Training set is generally used in order to building up a model.

Factors	J48
Correctly Classified Instances	39
Incorrectly Classified Instances	11
Accuracy	78 %

Table 1. Training set Batting First

Figure 2 showing the decision tree model for the bat-first where "venue" is the root of the tree. A decision tree is consists of decision node and leaf node. In this figure decision nodes are M6OW, M16ORN and the leaf nodes are win and lose.

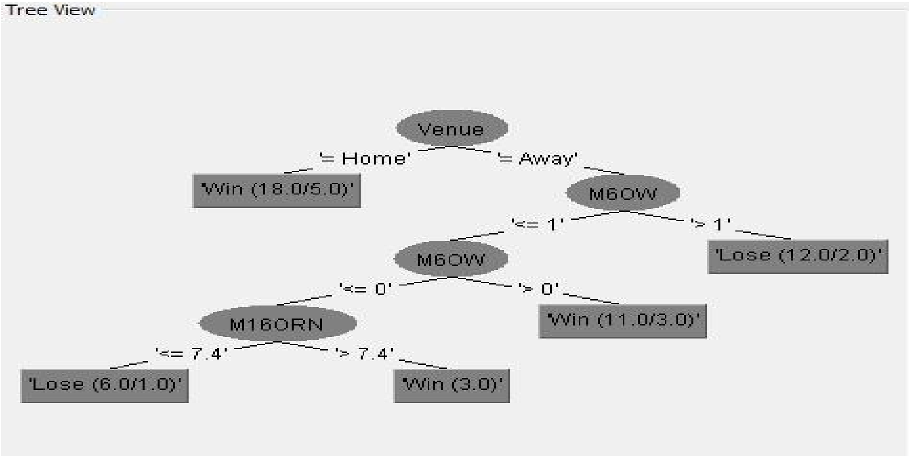


Fig. 6. Decision tree Bat-first

Evaluation of Test Set:

With the 70 % of the data used to create training set, we have another 30 % of the data to create the test set. Test set generally validate the model built by the training set. The accuracy of the test set showing 63.64 %.

Factors	J48
Correctly Classified Instances	7
Incorrectly Classified Instances	4
Accuracy	63.64 %

Table 2. Test set Batting First

Bat-Second In this part, the model showing the winning chance of CSK while batting second.

Evaluation of Training Set:

Below table describing the accuracy of the training set, while batting second. The accuracy of 82.5 % suggesting a better result than previous segment.

Factors	J48
Correctly Classified Instances	33
Incorrectly Classified Instances	7
Accuracy	82.5 %

Table 3. Training set batting Second

Test set of team batting second are showing 75 %.

Below figure 7 showing the Decision tree when the team batting second.

Factors	J48
Correctly Classified Instances	9
Incorrectly Classified Instances	3
Accuracy	75 %

Table 4. Test set batting Second

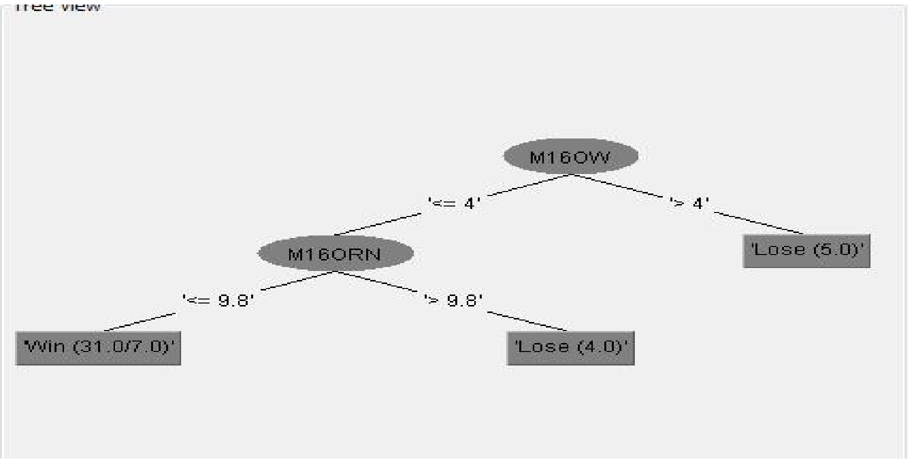


Fig. 7. Decision tree batting Second

In below table 5, we are showing the error comparison between bat-first and bat-second.

9 Prediction Modeling using Multiple Linear Regression

Multiple Linear Regression:

Error Comparison between 1st and 2nd segment		
Factors	Bat-First	Bat-Second
Kappa Statistics	0.5499	0.6067
Mean Absolute Error	0.3317	0.271
Root Mean Squared Error	0.4073	0.3681
Relative Absolute Error	56.3398 %	56.3398 %
Root Relative Squared Error	81.713 %	75.1307%

Table 5. Error Comparison between 1st and 2nd segment

Regression is an inherently statistical technique used regularly in data mining.Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. The multiple linear regression equation is as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p$$

where Y is the predicted or expected value of the dependent variable, X1 through Xp are p distinct independent or predictor variables, b0 is the value of Y when all of the independent variables (X1 through Xp) are equal to zero, and b1 through bp are the estimated regression coefficients. Each regression coefficient represents the change in Y relative to a one unit change in the respective independent variable. In the multiple regression situation, b1, for example, is the change in Y relative to a one unit change in X1, holding all other independent variables constant (i.e., when the remaining independent variables are held at the same value or are fixed). Again, statistical tests can be performed to assess whether each regression coefficient is significantly different from zero.

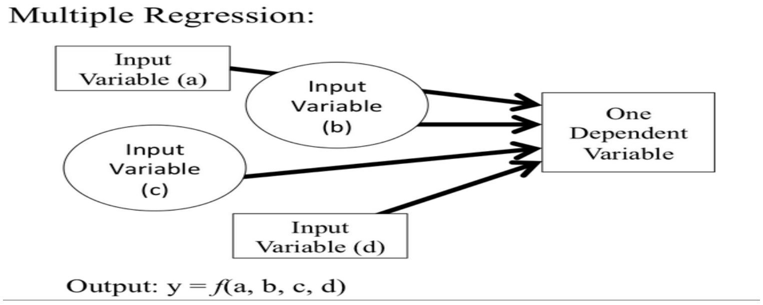


Fig. 8. Multiple Linear Regression

Formally, the model for multiple linear regression, given n observations, is

$$y_i = \beta_0 + \beta_p x_i 2 + ... + \beta_p x_i p f o r i = 1, 2, ...n.$$

In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. The least-squares

$$b_0 + b_1 + ... + b_p$$

estimates are usually computed by statistical software. The values fit by the equation

$$b_0 + b_1 x_i 1 + ... + b_p x_i p$$

are denoted y_i, and the residuals e_i are equal to y_i - y_i, the difference between the observed and fitted values. The sum of the residuals is equal to zero. The variance σ^2 may be estimated by

$$s^2 = \sum e_i^2 / n - p - 1$$

, also known as the mean-squared error (or MSE). The estimate of the standard error s is the square root of the MSE.

9.1 Multiple Regression Analysis Components:

Dependent variable (y): this is the variable representing the process you are trying to predict or understand (e.g., residential burglary, foreclosure, rainfall). In the regression equation, it appears on the left side of the equal sign. While you can use regression to predict the dependent variable, you always start with a set of known y values and use these to build (or to calibrate) the regression model. The known y values are often referred to as observed values.

Independent/Explanatory variables (X): these are the variables used to model or to predict the dependent variable values. In the regression equation, they appear on the right side of the equal sign and are often referred to as explanatory variables. We say that the dependent variable is a function of the explanatory variables. If you are interested in predicting annual purchases for a proposed store, you might include in your model explanatory variables representing the number of potential customers, distance to competition, store visibility, and local spending patterns, for example.

Regression coefficients (β): coefficients are computed by the regression tool. They are values, one for each explanatory variable, that represent the strength and type of relationship the explanatory variable has to the dependent variable. Suppose you are modeling fire frequency as a function of solar radiation, vegetation, precipitation and aspect. You might expect a positive relationship between fire frequency and solar radiation (the more sun, the more frequent the fire incidents). When the relationship is positive, the sign for the associated coefficient is also positive. You might expect a negative relationship between fire frequency and precipitation (places with more rain have fewer fires). Coefficients for negative relationships have negative signs. When the relationship is a strong one, the coefficient is large. Weak relationships are associated with coefficients near zero. β_0 is the regression intercept. It represents the expected value for the dependent variable if all of the independent variables are zero.

P-Values: Most regression methods perform a statistical test to compute a probability, called a p-value, for the coefficients associated with each independent variable. The null hypothesis for this statistical test states that a coefficient is not significantly different from zero (in other words, for all intents and purposes, the coefficient is zero and the associated explanatory variable is not helping your model). Small p-values reflect small probabilities, and suggest that the coefficient is, indeed, important to your model with a value that is significantly different from zero (the coefficient is NOT zero). You would say that a coefficient with a p value of 0.01, for example, is statistically significant at the 99% confidence level; the associated variable is an effective predictor. Variables with coefficients near zero do not help predict or model the dependent variable; they are almost always removed from the regression equation, unless there are strong theoretical reasons to keep them.

Residuals: These are the unexplained portion of the dependent variable, represented in the regression equation as the random error term, ϵ. Known values for the dependent variable are used to build and to calibrate the regression model. Using known values for the dependent variable (y) and known values for all of the explanatory variables (the X_s), the regression tool constructs an equation that will predict those known y values, as well as possible. The predicted values will rarely match the observed values exactly. The difference between the observed y values and the predicted y values are called the residuals. The magnitude of the residuals from a regression equation is one measure of model fit. Large residuals indicate poor model fit. Building a regression model is an iterative process that involves finding effective independent variables to explain the process you are trying to model/understand, then running the regression tool to determine which variables are effective predictor then removing/adding variables until you find the best model possible.

9.2 Problem Formulation:

We used same methodology as decision tree modeling while predicting with the multiple linear regression. A whole match divided into three segment and two different prediction for the bat-first and bat-second. Chennai Super Kings(CSK) used as the main team and prediction resulted CSK vs other teams. That means we predicted total runs and win possibilities of CSK against the other teams of the IPL. At first we predicted the total runs of three different segments when the team batting first and then the total runs of the three different segment when the team batting second.

10 Bat-First Prediction

By using our selected attributes we have discussed earlier that we predict the run scored by selected team when they batted first in segment Si, using the multiple linear regression algorithm. As there were no instantaneous attributes for the first segment run prediction so we just modeled a average score. In case of second segment run prediction we have chosen some specific attributes from first segment to give more weight. Finally in third segment we actually took the few strong attributes from first two segments and predicted the possible run can be scored.

10.1 First Segment

The prediction of first segment was quite difficult as there is no instantaneous attributes as we mentioned before. That is why we have just tried to find out some average run scored by the team who has batted first and the result was not so bad , the average run which we predicted using multiple linear regression it was just based on home or away attributes.For getting more close prediction we also add another attribute which is First Over Run(FOR).

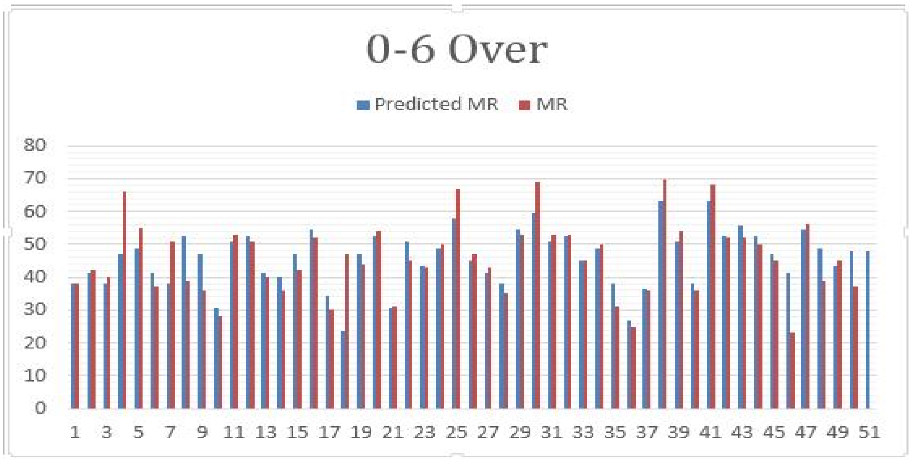


Fig. 9. First Segment run Prediction(Batting First)

10.2 Second Segment

In order to predict the second segment run we have analysed which attributes should take and which attribute will impact strongly. We have selected teams wicket, teams run rate, teams run and venue to give a prediction about probable run can be scored in second segment. Below the is a figure of our predicted model.

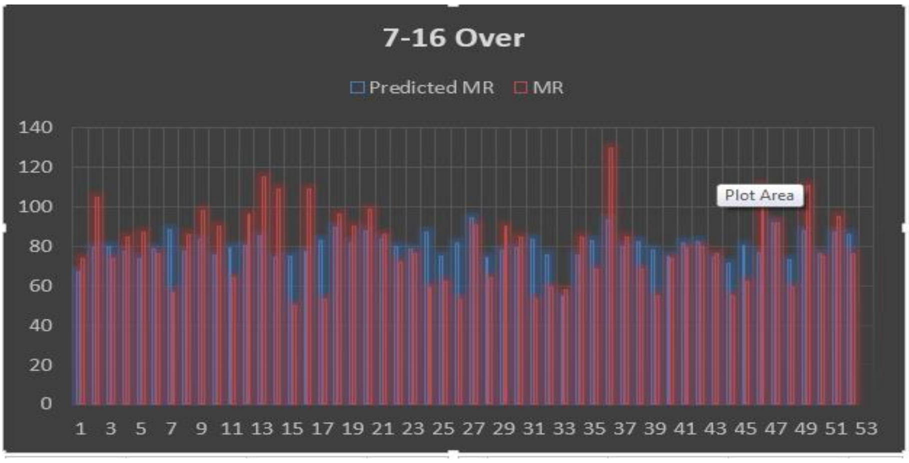


Fig. 10. Second Segment run Prediction(Batting First)

The coefficient values are interpreted as to how much of a unit change in Y will occur for a unit increase in a particular X predictor variable, given that the other variables are held constant. Here if we held Venue and MW constant then for MRN we would expect a -2.4 percent decrease on the Second Segment Run. So, we can see here the venue have positive coefficient as it has positive impact and other two attributes as negative impact on Second Segment Run prediction.

Second Segment Co-efficient	
Attributes	Coefficient
Intercept	101.4306
Venue	8.591546
MRN	-2.40124
MW	-4.71126

Table 6. Second Segment Co-efficient(Batting First)

As we have taken several key attributes for predicting the probable outcome of the second segment . We have got the attribute Home team wicket (MW) which we have collected from first segment has the most strong predictive value. It draws the main impact for predicting second segment outcome as you are viewing it in the table below.

Second Segment P-value	
Attributes	Coefficient
Intercept	$1.37 * 10^{-}9$
Venue	0.128311
MRN	0.132588
MW	0.083836

Table 7. Second Segment P-value(Batting First)

The regression line expresses the best prediction of the dependent variable (Y), given the independent variables (X). However, nature is rarely (if ever) perfectly predictable, and usually there is substantial variation of the observed points around the fitted regression line (as in the scatterplot shown below). The deviation of a particular point from the regression line (its predicted value) is called the residual value. Here the residuals is not that much marginal, in some cases the residual is very high because the unpredictable character type of cricket. Whatever, our final intention was building a model so that we can make a model with least residuals. That is where we are successful most of the cases.

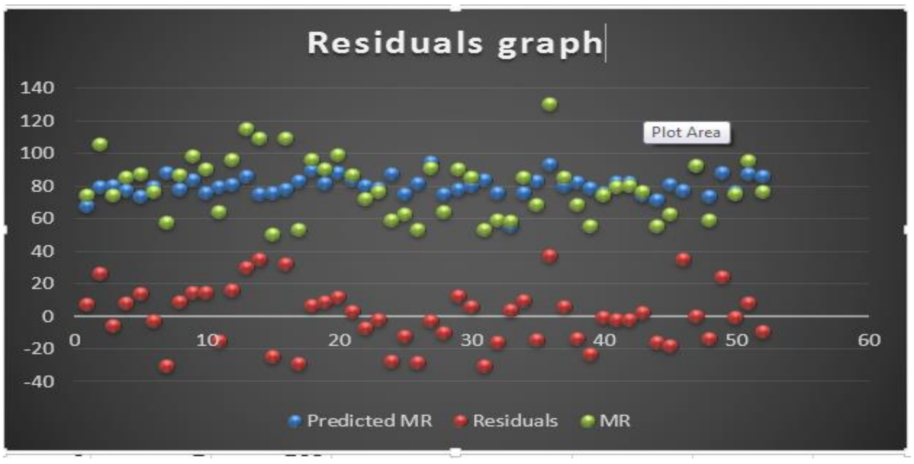


Fig. 11. Second Segment Residuals(Batting First)

10.3 Third Segment

Predicting third segment was really challenging as we wanted to predict the run a team can score in their final segment when the match is in progress. In the final segment of a T20 cricket, run rate tend to rise very much high as batsmen try to hit boundaries regularly to get a good total (In the first innings) or to achieve the target (In second innings). It is difficult to predict run in the final segment of the game because sometimes a lower order batsman can make some quick runs by slogging and the other day it may not be possible. So in the graph, it showing a very much disparity between predicted run and actual runs. We had to analyse and figure out which attribute should consider for third segment outcome prediction. we have used MRR,MW and Venue from segment 1 and segment 2.

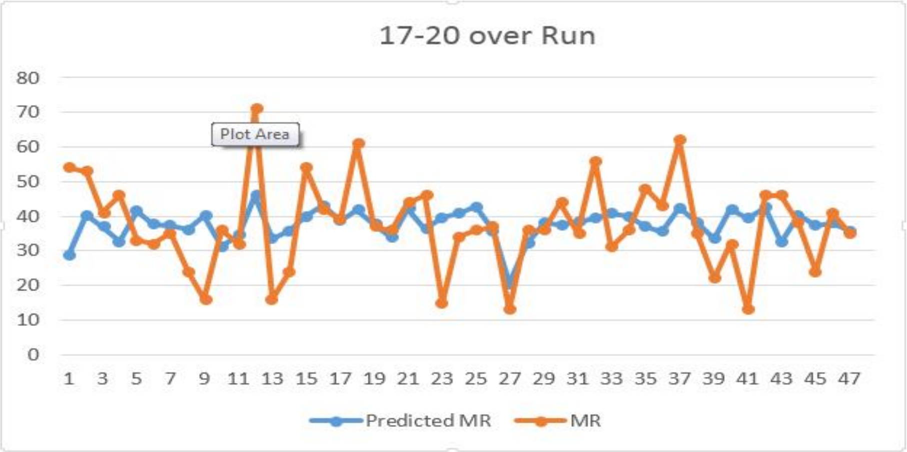
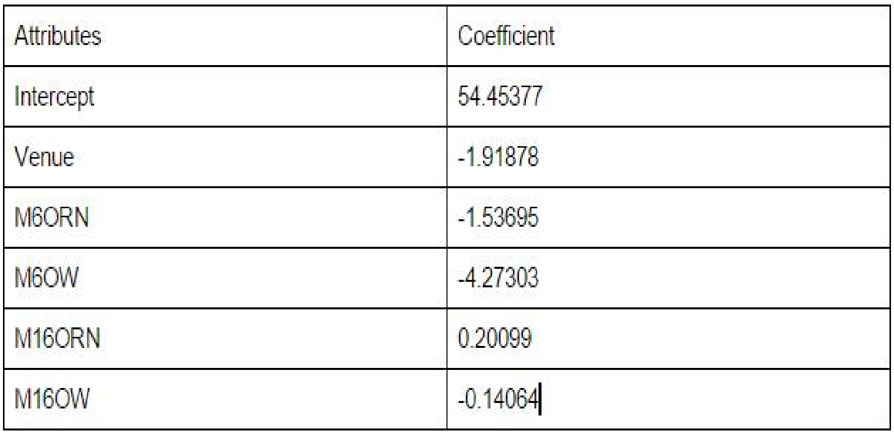


Fig. 12. Third Segment Run Prediction(Batting First)

Coefficient Table:

This are the coefficients of all attributes from Third Segment.



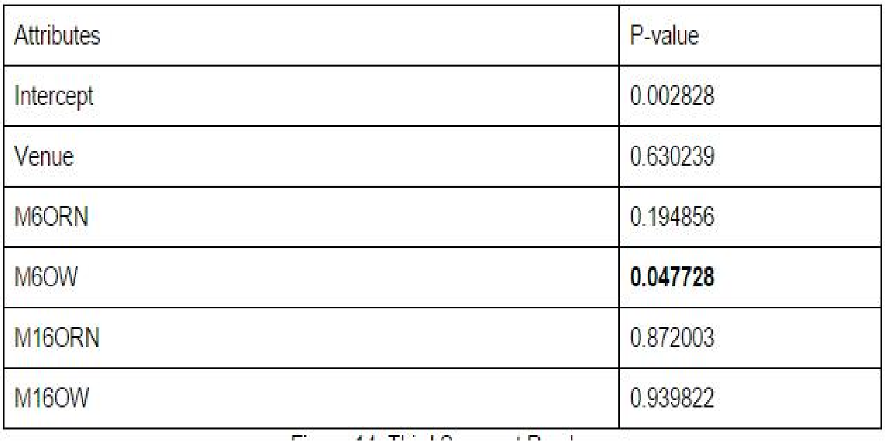
Attributes	Coefficient
Intercept	54.45377
Venue	-1.91878
M6ORN	-1.53695
M6OW	-4.27303
M16ORN	0.20099
M16OW	-0.14064

Fig. 13. Third Segment Coefficient(Batting First)

These are the Predictive values of all attributes from Third Segment.

11 Bat Second Run Prediction Model:

We have also created a linear regression model to predict the run when a team batting second. The prediction was done in three segments separately as the previous models. In this case we have made some modifications as when a team bat in a second innings there are some extra attributes we can consider for predicting the run. As this



Attributes	P-value
Intercept	0.002828
Venue	0.630239
M6ORN	0.194856
M6OW	**0.047728**
M16ORN	0.872003
M16OW	0.939822

Fig. 14. Third Segment P-value(Batting First)

is the second innings so we can get some instantaneous attributes for first segment which we didn't consider in the first segment of first innings.

11.1 First Segment

In that case we have selected few key attributes as the team is batting second and chasing run. We have taken the opponent team run rate and opponent team wicket along with venue attribute.This are the attributes which we didn't use for first innings first segment run prediction.So, it has given more accurate result.

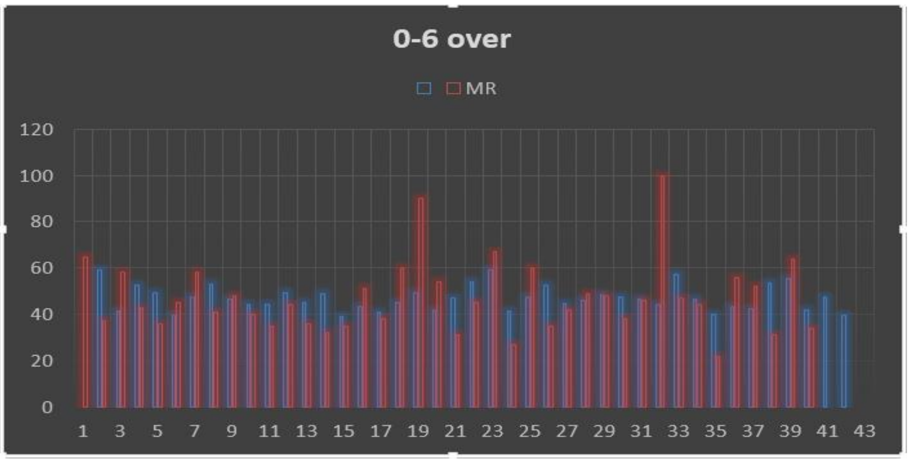
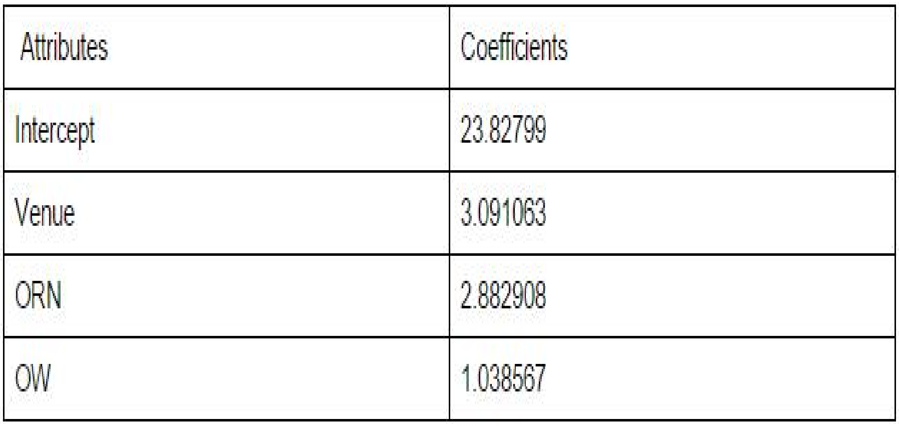


Fig. 15. First Segment run prediction(Batting Second)

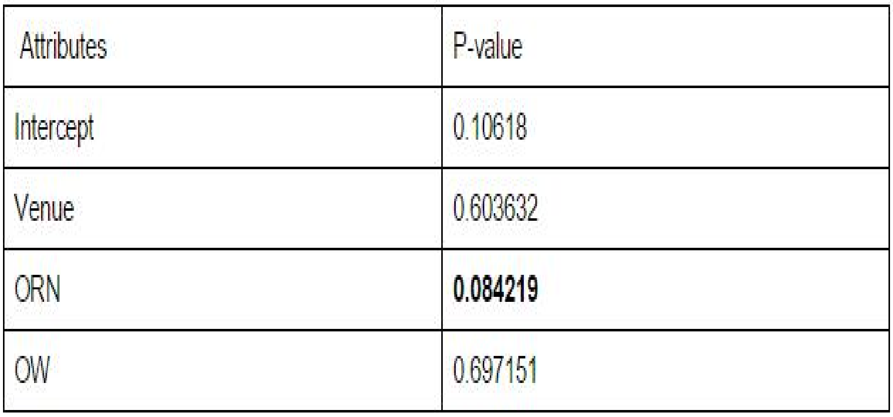


Attributes	Coefficients
Intercept	23.82799
Venue	3.091063
ORN	2.882908
OW	1.038567

Fig. 16. First Segment coefficient(Batting Second)

11.2 Second Segment

For that we have taken the first 6 overs my wicket and my run rate attribute and we have also taken the opponent teams wicket and run rate along with second segments run rate and wickets attributes.



Attributes	P-value
Intercept	0.10618
Venue	0.603632
ORN	**0.084219**
OW	0.697151

Fig. 17. First Segment P-Value(Batting Second)

This has given more accurate result than previous one. The residual is also very low as there is more attribute to make the model.

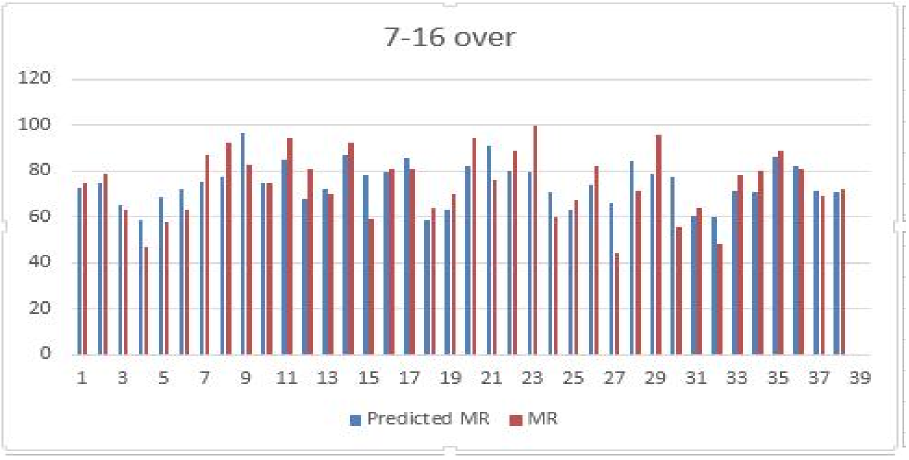
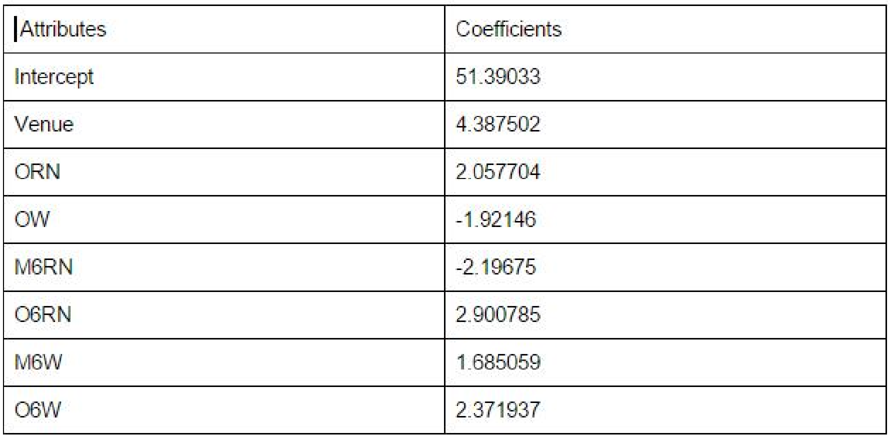


Fig. 18. Second Segment Run Prediction(Batting Second)

Coefficient:

These are the coefficient values of second segment attributes.



Attributes	Coefficients
Intercept	51.39033
Venue	4.387502
ORN	2.057704
OW	-1.92146
M6RN	-2.19675
O6RN	2.900785
M6W	1.685059
O6W	2.371937

Fig. 19. Second Segment Co-efficient(Batting Second)

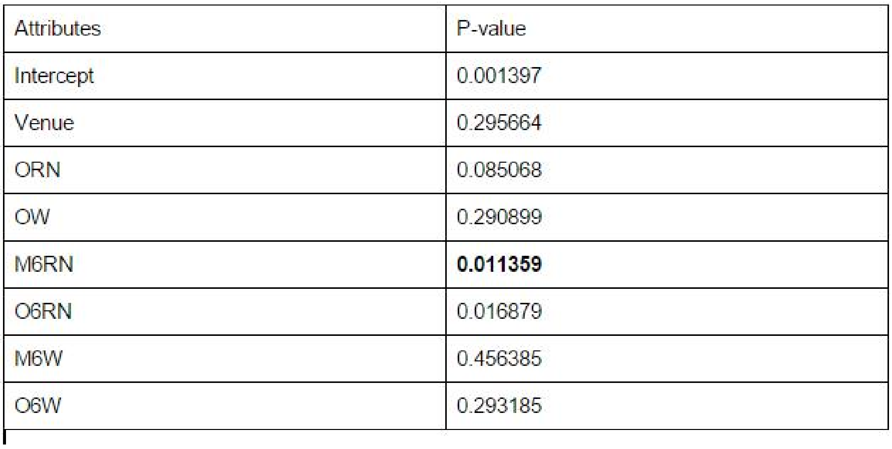
P-values:

These are the predictive values of attributes which have been considered for segment two.

Residuals of Second Segment:

11.3 Third Segment

Third segment prediction is the result of first and second segment prediction. Based on the result of first two segment, third segment prediction give its output.



Attributes	P-value
Intercept	0.001397
Venue	0.295664
ORN	0.085068
OW	0.290899
M6RN	**0.011359**
O6RN	0.016879
M6W	0.456385
O6W	0.293185

Fig. 20. Second Segment P-value(Batting Second)

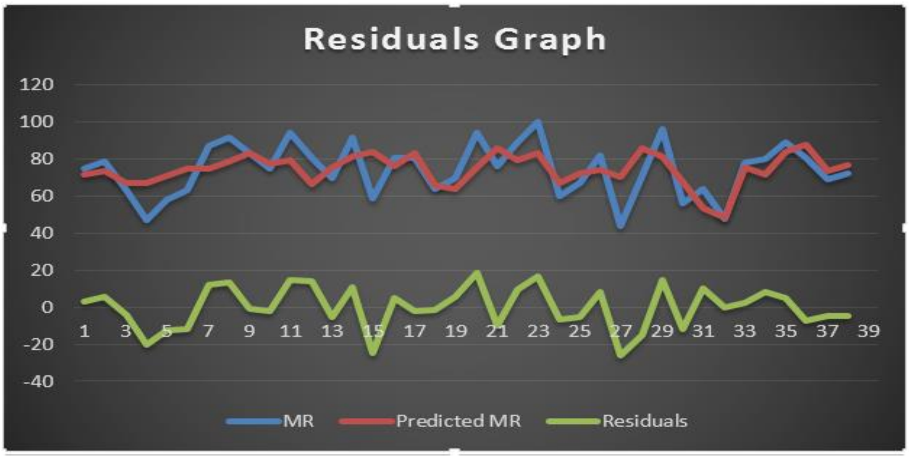


Fig. 21. Second Segment Residuals(Batting Second)

Coefficient:

These (Fig:23) are the coefficients values for all the attributes from Third Segment.

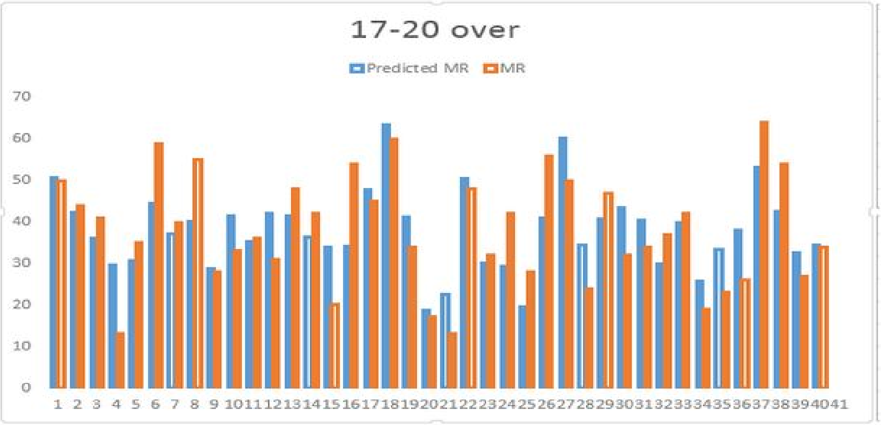
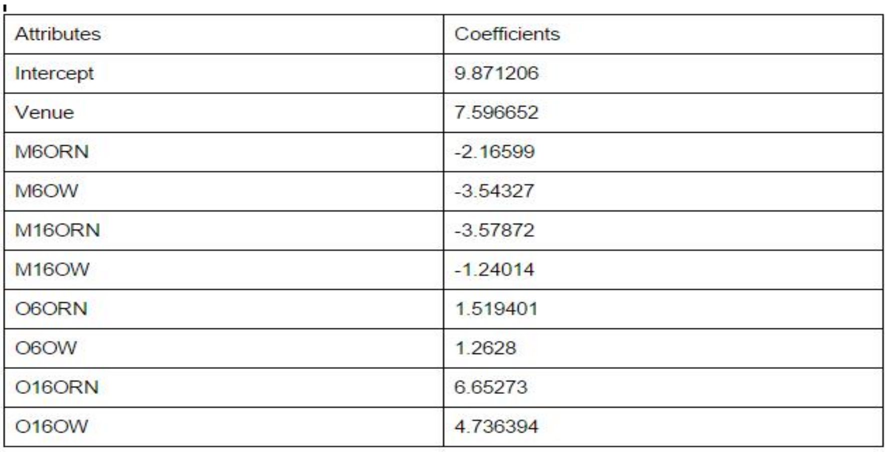


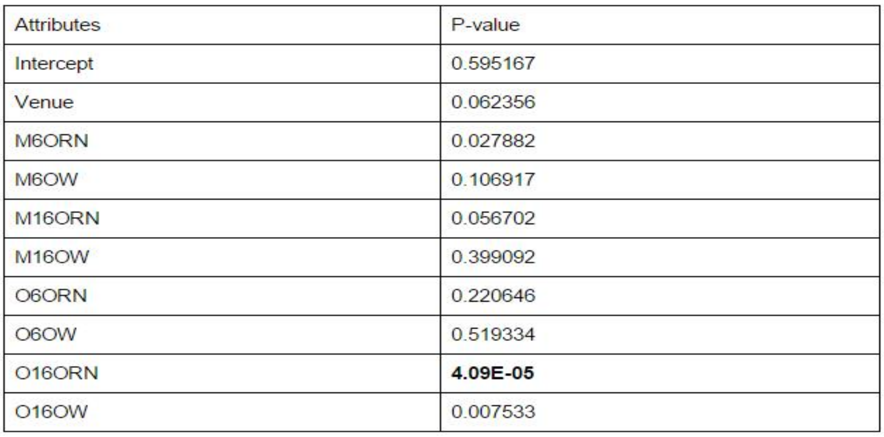
Fig. 22. Third Segment Run Prediction



Attributes	Coefficients
Intercept	9.871206
Venue	7.596652
M6ORN	-2.16599
M6OW	-3.54327
M16ORN	-3.57872
M16OW	-1.24014
O6ORN	1.519401
O6OW	1.2628
O16ORN	6.65273
O16OW	4.736394

Fig. 23. Third Segment Coefficient

P-value: These(Fig:24) are the p values for all the attributes from Third Segment.



Attributes	P-value
Intercept	0.595167
Venue	0.062356
M6ORN	0.027882
M6OW	0.106917
M16ORN	0.056702
M16OW	0.399092
O6ORN	0.220646
O6OW	0.519334
O16ORN	**4.09E-05**
O16OW	0.007533

Fig. 24. Third Segment P-Value

12 First and Second

12.1 Result Prediction based on First and Second Segment(bat first)

As we have divided our total model into three segment and we actually consider first two segment for predicting the match outcome as we wanted to find out the final match result when match is in progress. We have taken total 91 match for making our model using multiple linear regression and we have merged all the attributes from those matches based on different segment. After analyzing those two segment our model has given 75 % accuracy. So, we can predict any match outcome when the match is in progress based on our model. As we did not take any attributes from the team who will bat second and considering the attribute which we got from first segment, our predicted model is quite good. From the figure below we can see the graph view of our model, here 0 means lost and 1 means win. So, if the predictive final value is less than 0.5 then the result would be consider as lose and if the predictive value is greater than 0.5 then it would be consider as win. a

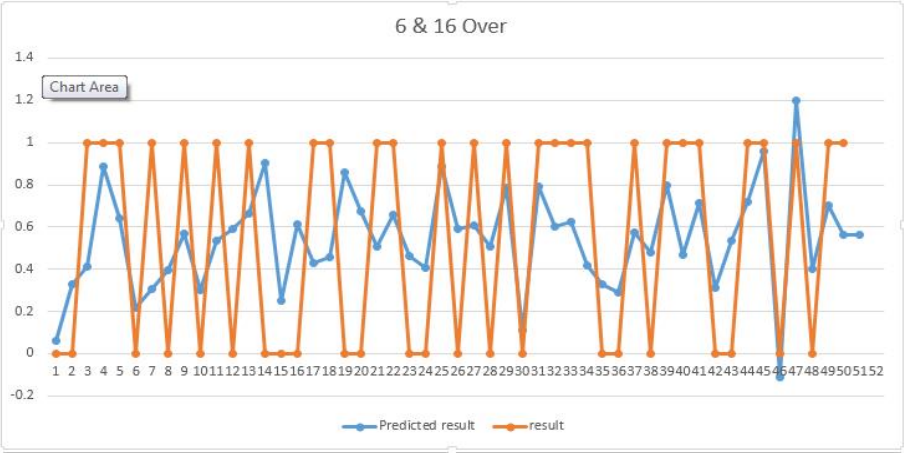


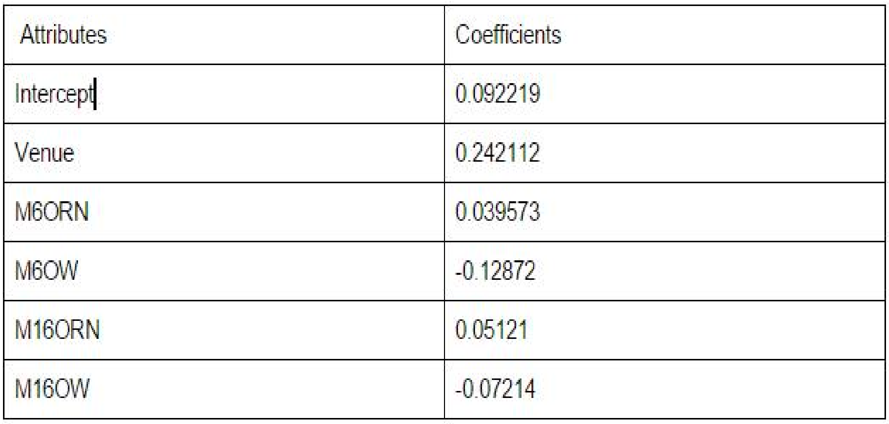
Fig. 25. First and Second segment prediction(bat first)

Coefficient: These (Fig-26)are the coefficients values for all the attributes from Win prediction based on bat first.

P-values: These (Fig-27) are the p values for all the attributes from Win prediction based on bat first.

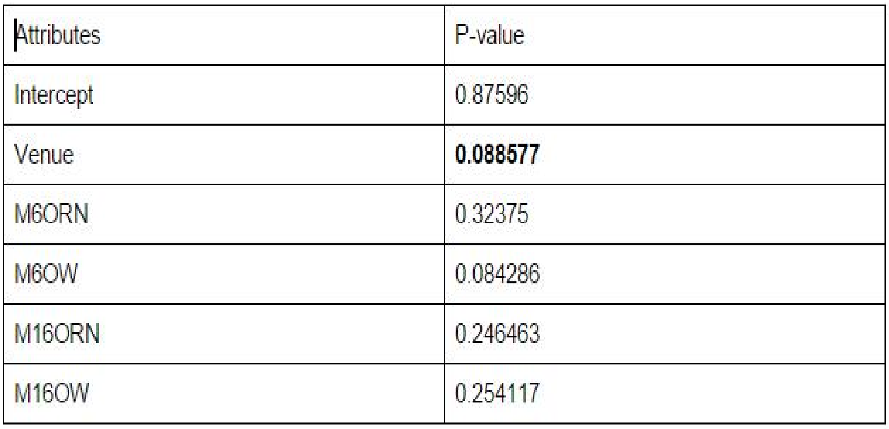
12.2 Result Prediction based on First and Second Segment(bat second)

:



Attributes	Coefficients
Intercept	0.092219
Venue	0.242112
M6ORN	0.039573
M6OW	-0.12872
M16ORN	0.05121
M16OW	-0.07214

Fig. 26. First and Second segment coefficient(bat first)



Attributes	P-value
Intercept	0.87596
Venue	**0.088577**
M6ORN	0.32375
M6OW	0.084286
M16ORN	0.246463
M16OW	0.254117

Fig. 27. First and Second segment P-value(bat first)

While calculating for 2nd innings segments we get the run rate value from team batting first. Which makes a better impact on a prediction model and that time our model has given 85.5 % accuracy which is really good.(Fig-28)

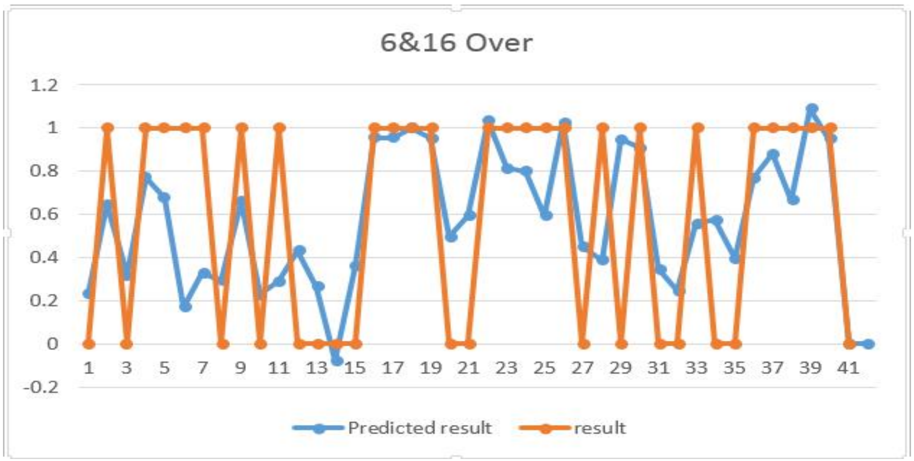
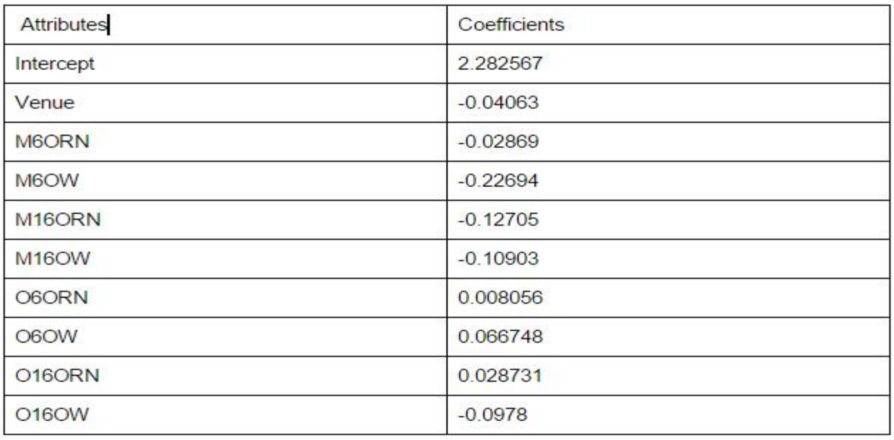


Fig. 28. First and Second segment Run-Prediction(bat second)

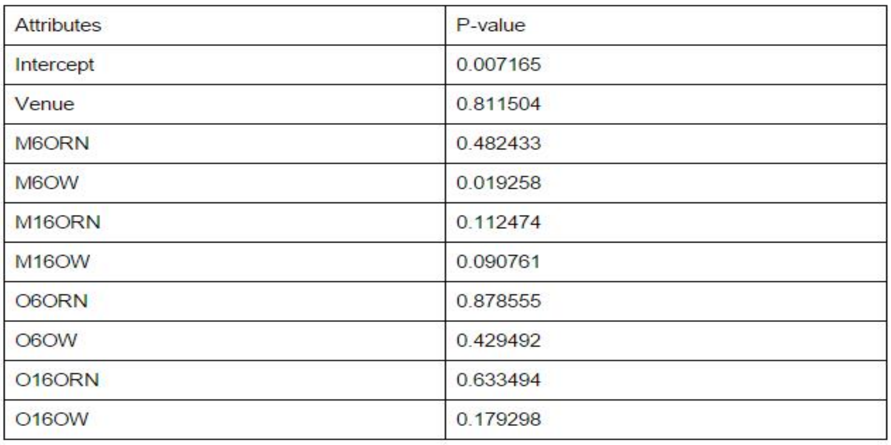
Coefficient: These (Fig.29) are the coefficients values for all the attributes from Win prediction based on bat second.



Attributes	Coefficients
Intercept	2.282567
Venue	-0.04063
M6ORN	-0.02869
M6OW	-0.22694
M16ORN	-0.12705
M16OW	-0.10903
O6ORN	0.008056
O6OW	0.066748
O16ORN	0.028731
O16OW	-0.0978

Fig. 29. First and Second Segment Coefficient(bat second)

P-values: These (Fig.30)are the p values for all the attributes from Win prediction based on bat second.



Attributes	P-value
Intercept	0.007165
Venue	0.811504
M6ORN	0.482433
M6OW	0.019258
M16ORN	0.112474
M16OW	0.090761
O6ORN	0.878555
O6OW	0.429492
O16ORN	0.633494
O16OW	0.179298

Fig. 30. First and Second Segment P-values(bat second)

13 Future Work

There are still a lot of space for improvement in our research. In future, we want to design a model which is much more efficient and give a very little error. Our model can not handle the result of the matches which are interrupted by rain or other natural calamities. We like to improve on that with the help of Duckworth-Lewis [5] method. In our research, we have divided a match in three segment in order to make our work easier. Our work in the future will be to design a model which can perform prediction in every over basis. We collected data manually by hand averaging the segment runs, wickets and run-rate. It was a very lengthy procedure and not efficient enough. So in future, we will work on to develop a web crawler which can crawl data according to our attributes selection and need.

14 Conclusion

Our main goal in this paper to develop a model to predict the outcome of a T20 cricket match while the game is in progress. We used the data of previous matches played between the team in order to design our model. We have used decision tree algorithm and weka to design this model. Efficiency and error checking were also done in our work. Along with decision tree, we also used multiple linear regression to predict match result with three segment. We have used two different technique to design our prediction model so we can easily find that which one is more efficient and giving lesser error. This knowledge will help us in the future to design a better prediction model.

References

[1] Sankaranarayanan VV, Sattar J. and Lakshmanan LVS. *Auto-play: A data mining approach to ODI cricket simulation and prediction.* In: Proceedings of the 2014 SIAM International Conference on Data Mining, 1064– 1072. SIAM 2014.

[2] Bhandari I, Colet E, Parker J, et al. *Advanced scout: Data mining and knowledge discovery in NBA data.* . Data Mining and Knowledge Discovery 1997; 1:121–125.

[3] Skinner B. *The problem of shot selection in basketball.* PloS one 2012; 7:e30776.

[4] Bailey M, and Clarke SR. *Predicting the match outcome in one day international cricket matches, while the game is in progress.* Journal of Sports Science and Medicine 2006; 5:480.

[5] Duckworth, F.,and T. Lewis *Your Comprehensive Guide To The Duckworth/Lewis Method For Resetting Targets In One-Day Cricket* University of the West of England,1999.

[6] Theja Tulabandhula and Cynthia Rudin. *Tire Changes, Fresh Air, And Yellow Flags: Challenges in Predictive Analytics For Professional Racing;* Massachusetts Institute of Technology, Cambridge, Massachusetts, June 2014.

[7] S. Akhtar and P. Scarf. *Forecasting test cricket match outcomes in play* Salford Business School, University of Salford, 2011.

[8] D. Roy Choudhury, Preeti Bhargava, Reena, Samta Kain; *Use of Artificial Neural Networks for Predicting the Outcome of Cricket*

Tournaments Department of Computer Engineering, Delhi College of Engineering, New Delhi, 2007.

[9] Ganeshapillai G, Guttag J. *A data-driven method for ingame decision making in MLB: When to pull a starting pitcher.* In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 973–979. ACM 2013.

[10] Rupali Bhardwaj and Sonia Vatta; *Implementation of ID3 Algorithm* International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013

[11] J.R.Quinlan; *Induction of Decision Trees* Centre for Advanced Computing Sciences, New South Wales Institute of Technology, Sydney 2007, Australia