National College of Ireland

# Analysis of Batsman Contribution using Classification Algorithm : A Limited Over Cricket Perspective

MSc Research Project
Data Analytics


Harish Malli Dhanarajan
x15008991

School of Computing
National College of Ireland


Supervisor: Dr.Eugene O'Loughlin

**National College of Ireland**
**Project Submission Sheet – 2016/2017**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Harish Malli Dhanarajan |
| **Student ID:** | x15008991 |
| **Programme:** | Master of Science in Data Analytics |
| **Year:** | 2017 |
| **Module:** | Research Project |
| **Lecturer:** | Dr.Eugene O'Loughlin |
| **Submission Due Date:** | 08/05/2017 |
| **Project Title:** | Analysis of Batsman Contribution using Classification Algorithm : A Limited Over Cricket Perspective |
| **Word Count:** | 5389 |

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:** | 8th May 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): | |

# Analysis of Batsman Contribution using Classification Algorithm : A Limited Over Cricket Perspective

Harish Malli Dhanarajan
x15008991
MSc Research Project in Data Analytics

8th May 2017

**Abstract**

Limited over cricket basically brings entertainment to the people and they like it since the matches played will be finished in shorter time with a full pace of cricketing shots involved. So the history of Test cricket has got faded away nowadays. Making use of their growth, some of the companies started to gradually change the game of cricket into business. One such business platform led to the evolution of Indian Premier League (IPL). There are 8 teams in an IPL league and each team has a manager who selects the player by buying him through bidding. This research is an attempt to help the managers to select the right player by estimating best price that he deserves. In order to achieve this, a model was developed for predicting the contribution that a player can make in fore coming matches based on previous performance details. Initially, data pre-processing is done and algorithms that perform feature selection such as Boruta package, *fscaret* library are used to predict the important variables in order to maintain higher accuracy for the model. Followed by this, Decision tree and Naive Bayes classification algorithm is performed to predict positive and negative contributor. Model evaluation metrics such as accuracy, specificity, sensitivity and kappa value showed that Decision tree was the best fit model. Finally as a completion of this research two case studies are presented.

## 1 Introduction

In general, data mining is a technique in which extraction of knowledge takes place from a large set of databases that is popularly known as Big data. Retrieving knowledge and information from a complex data is difficult for any human to deal with, but data mining tools performs them with ease. The advent of technology has led a way to capture what is happening and convert them into data. In this fast moving game of cricket, test matches are being faded out and limited over cricket are growing more and more among the people. One such limited over tournament is the Indian Premier League where players are selected based on previous performance and certain amount of money is paid to them for representing a particular team. Selection of required player is done by bidding process which is anyway not relevant to this research. Basic prediction of a player performance begins with the evolution of Duckworth Lewis model as described in the research done by

(Beaudoin and Swartz; 2003). Apart from this basic research, there are so many methods used by various researchers in prediction of a player performance. One of the namely research is (Barr and Kantor; 2004) which makes use of a two dimensional approach using strike rate and average as a selecting criterion for predicting a batsmen performance. By reviewing most of researches, the variables that are repeatedly used for predicting player performance in limited over cricket are average, strike rate and boundaries hit by a batsmen.

The overall objective of this research is to develop a model for predicting a batsmen contribution and then a small attempt is taken to compare the batsmen contribution with the price paid to them during an auction. The motive is to help the managers to easily consider on which batsman deserves how much money for auctioning. Some of the key hypothesis that are about to be achieved are as given below,

- Does a highly averaged batsman who is capable of scoring more than 20 runs per innings necessarily be a positive contributor of a team?
- Is the price value of a batsman justified depending on his previous record or consistency of the game?

The structure of this research project is as follows: section 2 consists of related works done in accordance of predicting the player performance in limited over cricket ; section 3 comprises a discussion of methods to be implemented and model evaluation metric; section 4 is about the dataset description, data preparation and implementation of the model and section 5 consists of evaluation of the implemented model with two case studies, then followed by conclusion and future works.

# 2 Related Work

This section describes the research and works done by other authors that includes an overview of concepts used in cricket, existing theories related to the research and inconsistency or limitations of the existing research. Other ideas discussed are benefits and goal of this study.

## 2.1 Concept of Cricket

Cricket is a bat and ball game played head-to-head consisting of eleven players on each side along with a substitute player. Rather than discussing rules of the game in this section, some of the statistics used in cricket are discussed by the author. Some of the statistical concepts involved in batting are batting average and strike rate. (Beaudoin and Swartz; 2003) defines the average as fraction of total runs to that of number of matches the batsmen was out. Secondly, they discusses about the strike rate as total number of runs scored to that of number of balls faced. In the next section, some of the existing theory based on team selection and prediction of match outcome.

## 2.2 Player Performance in All Forms of Cricket

The process of selecting a team in limited over domestic cricket is entirely different from International Cricket. A player is bid for the best price and taken into the team based

on factors like previous performance, fairness of play and fitness level. Researchers shows concentrations on cricket nowadays for predicting a player performances.

In general, the analysis of player performance can be illustrated using the work of (Beaudoin and Swartz; 2003). They have made use of batting average, strike rate and runs scored per innings for calculating the performance of a player. Simple statistics that yields standard error was considered and research was carried out. But they also admit that data collections were tough and statistical interpretation did not show better result. To overcome this problem, a research done by Maheswari and Raman(2009) can be used. Data collection was made using match videos and was stored in an object relational database since the data generated was complex in nature. Then, Principal Component Analysis was carried out as a part of dimensionality reduction and important variables were determined. As a part of data mining, Association Rule was applied that calculates frequent patterns among the data based on their likelihood (Ma and Liu; 1998) and the output was interpreted. The output of Maheswari and Raman (2009) describes how a player has faced a bowling attack and also determines what is the likelihood of how a batsmen gets out. Although, the algorithm and output were outstanding, this same set of analysis can be carried out on only less than 50 matches. Then in 2011, (Scarf et al.; 2011) has proposed a research to predict the player performance in test cricket. Since test cricket consists of four innings, Scarf et al(2011) has made use of negative binomial distribution and logistic regression to predict the score that a batsmen can make in third and fourth innings based run rate of a batsmen in first two innings. But, this model was considered to be covncentrating only on run rate and not on other factors. Yet another statistical analysis was carried by (Wickramasinghe; 2014) based on hierarchy model. It basically comprises three levels and a mixed level. At each level, separate analysis was carried out based on average, player body measurement and runs scored. Finally, a prediction was made on batsmen score. Again, as like above research made by Scarf et al(2011) and Maheswari and Raman (2009), (Wickramasinghe; 2014) has also faced a limitation in the research such as usage of weak parameter for analysis and only limited player scores were predicted. (Lemmer; 2008) has come up with a calculation to assign rank for each player but fails to explain evaluation of the results briefly. The factors considered were average, strike rate and runs scored by a batsmen. But the calculation were performed on less amount of players.

The availability of data was more after the advent of internet. Hence to overcome all the research problems described above, several advanced form of data collection and analyses were carried out that will help yielding better output compared to statistical models. (Sankaranarayanan et al.; 2014) made use of ridge regression, attribute bagging (Ensemble model on random features) and Nearest Neighbour clustering. Data related to batting average, bowlers details and how batsmen gets out were collected for 125 matches. A ten-fold cross validation was performed on the dataset before analysis. An accuracy of about 70 percent was shown on all the models. Sarkar and Banerjee (2016) also predicted the player performance based on average, strike rate, number of boundaries, opposition team, dismissal and batting position for all the players who have scored above 8000 runs. The methodology carried out were Exponential Distribution, Maholanobis Distribution, Maximum Likelihood Estimate and Weibull Distribution. They also concludes that Weibull distribution has shown a good fit to the real time data when compared to exponential distribution and other traditional methods. He also concludes that player rating can be determined using Maholanobis distribution. (Iyer and Sharda; 2009) determined how players perform during world cup tournaments. The methodology

used was all types of neural network such as Multi-Layer preceptron, Linear Model and Radial basis function network. All these model have shown accuracy of over 75 percent in prediction. As like all the other research, Iyer and Sharda (2009) also used average, strike rate, dismissal and number of boundaries secured. Finally, it has been made clear that most of the work done on predicting player performance are based on batting average, strike rate and runs secured by a batsmen. Out of all these researches neural network produced by Iyer and Sharda (2009) has shown greater value of accuracy in prediction player performance.

## 2.3 Works on Indian Premier League

At first, (Ahmed et al.; 2013) has made a research based on how to select valid players into a team with a finite budget. A multi criterion decision making strategy was carried out by them . Variables considered were batting average and bowling average in order to come up with the decision model. Another approach that resembles ahmed2013multi work is the research done by (Barr and Kantor; 2004). Barr and Kantor (2004) developed a selection criterion for comparing and picking up the right batsmen in limited over cricket. The selection criterion was based on probability of a batsmen getting out with respect the strike rate and average of a batsmen. These theoretical approach was practically implemented by (Kansal et al.; 2014) for evaluating the players and also for deciding how much money a player should be paid. This research has made use of all the important variables such as batting average, strike rate, number of century, number of boundary and so on. A conclusion was brought out by implementing classification models such as decision tree, Naive Bayes and Multilayer Perception algorithm. The dependent variable was the price paid for the player based on the performance. The drawback of Kansal et al(2014) work was due to the over fitting of data that resulted in inappropriate results.

## 2.4 Performance Prediction in Other field of Sports

In football, Poison distribution was used to determined the outcome of a match in most of the research and one of the main research was (Dyte and Clarke; 2000). It has considered venue, rating of a team and goals scored as major factors. This model was considered one of best in the field of research in football. Another most commonly used technique is the Bayesian approach by (Suzuki et al.; 2010) in order to determine the classification probability that leads in predicting the outcome of a match. Apart from football, Zimmermann et al(2013) and Loeffeelholz et al(2013) have proposed data analysis model on Basketball match prediction. (Zimmermann et al.; 2013) has considered Machine learning techniques such as decision tree, Naive Bayes, Random Forest and Artificial Neural Network. A clear view is not presented on which is the accurate model. (Loeffelholz et al.; 2009) also came up with a model that uses feed-forward neural networks, probabilistic neural network and generalized regression neural networks. In an overall research, Feed forward Neural Network has given good accuracy on all the moves such as shooting the ball, passing and goal results.

# 3 Methodology

The methodology chosen to guide this project is CRISP-DM. It generally consists of six stages namely: 1. Business Understanding, 2. Data Understanding, 3. Data Preparation,

4. Modelling, 5. Evaluation and 6. Deployment (Chapman et al.; 2000). These six steps are often performed as an iterative process in order to keep the cycle moving for a business.

## 3.1 Software Used

The list of software used in this research are as follows,

- R studio 64 bit (Version 3.3.1) for Web Scrapping and Modelling
- MS Excel
- Google Refine
- Tableau for data visualisations

## 3.2 Model Evaluation Terminologies

The most common evaluations used for interpreting the accuracy is Sensitivity, Specificity and Kappa Value. Sensitivity is the capability of a test dataset to correctly classify instances is positive and Specificity is defined as the correctly identified instances that are not positive. Kappa value is used to determine on how good a model is fit for the data. More the kappa value, greater the goodness of the fit. The formula is given below,

- Sensitivity : (No.of true positive) / (No.of true positive + No. of false negative)
- Specificity : (No.of true negative) / (No.of true negative + No. of false positive)

## 3.3 Methods Used

Methods used in this research were feature selection, decision tree and Naive Bayes classifier algorithm. Variable selection or feature selection is a technique to select the important variable that are required for the analysis. This can result in increase in model accuracy and helps the model to perform faster as well. Several other advantage of performing feature selection is that the noise reduction can takes place among the variables (Guyon and Elisseeff; 2003). There are two main types of feature selection methods, such as wrapper and embedded method. In this research, wrapper method is used for selecting the features. This method involves searching of best feature by using induction algorithm which creates appropriate classifiers that is required for the feature set (Kohavi and John; 1997). Then, highly evaluated classifiers are taken into consideration and result is presented. The wrapper method used in this research are performed using Boruta algorithm and fscaret package.

Decision tree that is used in the research is the Conditional Inference tree. The choice of using this tree is because it works well on continuous, ordinal and also for multivariate response variables (Hothorn et al.; 2015). In the implementation section, the test statistics [1] was used to compute the output class (Positive or Negative). The tree split is done if the p value is less than 0.05. (Knijnenburg et al.; 2009) states that the appropriate permutation test[2] values are determined similar to the hypothesis

[1]Statistical value used to perform the hypothesis test

[2]Test to determine the sampling distribution of t statistics

derived from test statistics and then the tree is formed. Another advantage of Conditional Inference tree is that, it avoids over fitting of the dataset.

Followed by decision tree, the Naive Bayes classifier algorithm is modelled. The package used for Naive Bayes classifier was klaR that performs functions such as classification and visualization. It is a probabilistic classifier model that is designed using a group of algorithm. For example, a ball is said to be cricket ball if it is round, red colour and has stitches in them. It generally works based on Bayes Theorem (Zhang; 2016) that is usually expressed as P(A Given B) = P(B Given A)*P(A)/P(B).

The justification for choosing Naive Bayes Model for the collected dataset was because the data satisfies one of the Bayes theorem as in (Pawlak; 2002). The theorem states that *The Naive Bayes classifier is optimal for any two-class concept with nominal features that assigns class 0 to exactly one example, and class 1 to the other examples, with probability 1.* Furthermore, Naive Bayes algorithm is said to avoid over-fitting. All these Naive Bayes theory is applied using an algorithm in the implementation section.

# 4 Implementation

This section consists of dataset collection, data pre-processing, feature selection and predictive models.

## 4.1 Dataset Collection

There are two dataset used in this research. The first data is carried out throughout the project whereas the second dataset is used only in case study 2 as part of comparing price between positive and negative contributors. These datasets are obtained using web scraping by R studio.

- The first dataset was taken from www.bigbashboard.com. This data consist about the entire player details who represented in limited over matches (20 Overs). It also comprises of every key information required to estimate a batsmans performance in his career. The target variable is *Effectiveness*. This dataset is represented as Dataset 1 throughout the research paper. The description of this dataset is given in appendix A.

- The second dataset was from the website, www.iplt20.com. This dataset is about Indian Premier League which is one of the tournaments played in limited over matches (20 Overs). It comprises information like origin of the player, who bought the player and price paid for the player. This dataset is only used for a case study and it is represented as Dataset 2. The description of this dataset is given in appendix B.

## 4.2 Data Pre-Processing

This sub-section consists of dataset pruning, handling with missing values and data transformations as required for the analysis.

### 4.2.1 Data Pruning

Dataset pruning is generally referred to cutting away unwanted data that are not required for the analysis. The pruning was done on both the dataset in order to get rid of outlier as well as under-performed player that might affect the final modelling result. Some of the variable that are pruned are given below,

- Match (From Dataset 1) : Those who played less than 10 matches were removed since they do not have enough information for predicting their contribution.
- Total Runs (From Dataset 1) : On an average a batsman should score at least 20 runs in an innings. The threshold of 20 runs was derived from a separate dataset that consists of ball by ball score of each batsman. So, Batsmen below 200 runs (which is 20 Runs * 10 Match) was removed.
- Type (From Dataset 2) : Bowlers are removed since this analysis is only for Batsmen, All-Rounder and Wicketkeeper (who basically comes under Batsmen category).

### 4.2.2 Handling Missing Values

Once dataset pruning is done, the missing values are detected and analyzed for their absence. The missing values in both the datasets were not because of the wrongly entered or mistakenly avoided data value. Rather it was due to the fact that the batsman has not performed that specific task. For eg. 100s has so many missing values because scoring a century was difficult in limited over matches. Hence missing values were replaced with Zero.

### 4.2.3 Data Transformation

Variables that were transformed in the first dataset were Average, Strike Rate, Effectiveness (Target Variable) and in the second dataset was Price Paid (Target Variable).

- Average was normalized as below 20 (As described in subsection 4.2.1), 20 to 25, 25 to 30 and Above 30 (Very few Batsman have this).
- Strike Rate was normalized into Below 100 (Batsmen score 6 run in 6 balls), 100 to 120 (Batsmen score 6 runs in 5 balls), 120 to 150 (Batsmen score 6 runs in 4 balls) and above 150 (Batsmen who score 6 runs in less than 3 balls).
- Effectiveness (Target Variable) was categorized into Positive (if above 0) and Negative (if below 0).
- Price Paid (Target Variable) was converted into Million since it was in Lakhs (Indian Notation).

The transformation of Average, Strike Rate did not affect the final result. It helped to categorize a batsmen easily according to their Average and Strike Rate.

## 4.3 Feature Selection

Comparing both the datasets, the first dataset had many features. As per cricket theory, all the variables in dataset 1 can be considered important. So the author has performed feature selections to determine which variable to exactly chose for better output and prediction result. It was performed using Wrapper Algorithm such as Boruta Algorithm, Gradient Boosting and Principal component analysis with Neural Network (pcaNNet). The reason for choosing wrapper algorithm is because of their stability and speed of calculation. The package used for feature selection was *fscaret* and Boruta. The feature importances are calculated according to the generalization error and the induction algorithm. *fscaret* library consists of 227 packages with about 129 classification models in which Gradient Boosting (GBM) and Principal component analysis with Neural Network were one of them.

### 4.3.1 Feature Selection with Boruta Algorithm

Boruta package consists of a wrapper algorithm that performs feature selection using Random Forest based on a dependent variable *Effectiveness*. In this, a shadow feature is created first and Random Forest Classifier was applied to it with several iterations until the decision was made. As a result, out of 30 variables only 1 was considered to be rejected as shown in figure 1. Basically the variable rejection is done if there exists a negative mean importance and median importance. It shows that, only player varible was rejected. Then, The plot of feature importance is shown in figure 2 and top 10 variables were selected for analysis.

| | meanImp | medianImp | minImp | maxImp | normHits | decision |
|---|---|---|---|---|---|---|
| Player | -0.211359 | -0.07550552 | -1.5064929 | 1.031867 | 0.0000000 | Rejected |
| Match | 5.510904 | 5.74435360 | 4.1057529 | 6.714088 | 1.0000000 | Confirmed |
| Innings | 5.346179 | 5.21392491 | 3.9513025 | 6.604863 | 1.0000000 | Confirmed |
| Not.Out | 5.820608 | 5.76636154 | 3.9546578 | 8.002002 | 1.0000000 | Confirmed |
| Batting.Run | 8.130952 | 8.12653081 | 7.0287765 | 9.128914 | 1.0000000 | Confirmed |
| High.Score | 11.530230 | 11.76756277 | 9.1689736 | 12.911331 | 1.0000000 | Confirmed |
| Batting.Average | 17.748683 | 17.61751273 | 16.8177316 | 19.064572 | 1.0000000 | Confirmed |
| Balls.Played | 7.148160 | 7.30184220 | 5.4713872 | 8.359942 | 1.0000000 | Confirmed |
| Strike.Rate | 9.093083 | 9.13124236 | 8.0440289 | 10.691411 | 1.0000000 | Confirmed |
| Hundread | 4.793157 | 4.75715236 | 3.8077251 | 5.874284 | 1.0000000 | Confirmed |
| Fifty | 12.547298 | 12.68765110 | 11.1077994 | 13.446333 | 1.0000000 | Confirmed |
| Thirty | 5.706070 | 5.76346594 | 4.8066024 | 6.714422 | 1.0000000 | Confirmed |
| Ducks | 3.171588 | 3.46530925 | 0.7677418 | 4.270247 | 0.8695652 | Confirmed |
| Four | 9.487805 | 9.45482062 | 8.4695440 | 10.836373 | 1.0000000 | Confirmed |
| Six | 7.541265 | 7.51054823 | 6.9425528 | 8.866667 | 1.0000000 | Confirmed |
| BpB | 13.634969 | 13.74616953 | 12.4917562 | 14.457752 | 1.0000000 | Confirmed |
| Bp4 | 10.606302 | 10.64257815 | 9.5033162 | 11.757785 | 1.0000000 | Confirmed |
| Bp6 | 5.560564 | 5.49547804 | 4.2652220 | 7.619985 | 1.0000000 | Confirmed |
| PrB | 8.618277 | 8.65530452 | 7.1064308 | 9.973977 | 1.0000000 | Confirmed |
| Pr4 | 5.796179 | 5.98096409 | 3.4259467 | 7.414177 | 1.0000000 | Confirmed |
| Pr6 | 4.023144 | 3.76701642 | 2.5826241 | 6.477703 | 0.9565217 | Confirmed |
| BBP | 13.640057 | 13.69876416 | 12.8268564 | 14.936650 | 1.0000000 | Confirmed |
| RpB | 3.516590 | 3.64501791 | 2.3279945 | 4.644845 | 0.9565217 | Confirmed |
| BIx | 12.873666 | 12.96194216 | 11.9674427 | 13.920860 | 1.0000000 | Confirmed |
| RpNB | 3.503062 | 3.35468436 | 2.0678676 | 4.474923 | 0.8695652 | Confirmed |
| ScIx | 14.685577 | 14.82063060 | 13.6831011 | 15.569908 | 1.0000000 | Confirmed |
| RpI | 35.441598 | 35.36252606 | 33.7187853 | 37.544654 | 1.0000000 | Confirmed |
| BpI | 18.882177 | 18.78416843 | 17.6070064 | 20.199394 | 1.0000000 | Confirmed |
| ABP | 9.876991 | 9.99575659 | 8.7271727 | 10.669459 | 1.0000000 | Confirmed |
| Basra | 24.862446 | 24.69792817 | 23.7151559 | 26.487588 | 1.0000000 | Confirmed |

Figure 1: Decision of Boruta Algorithm

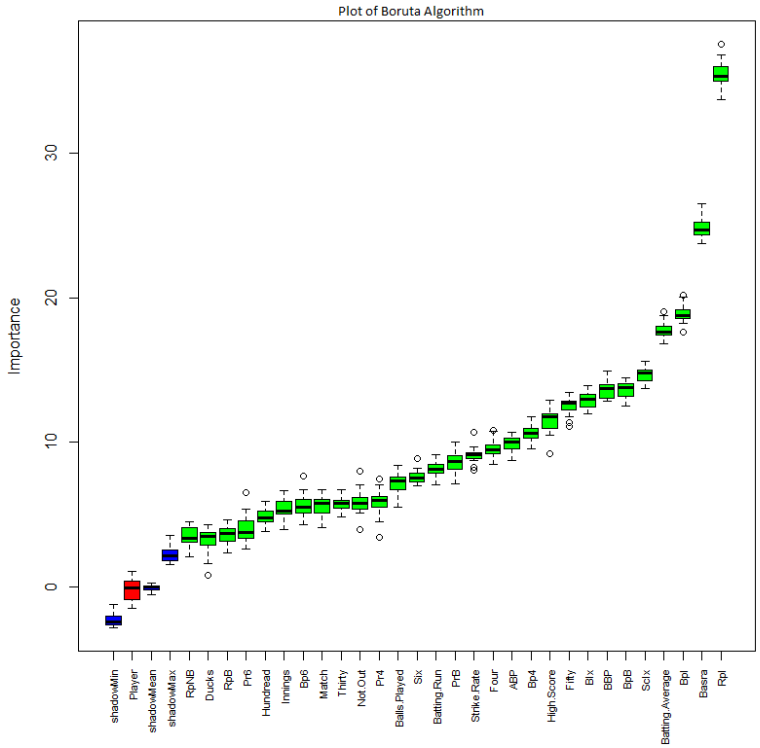


Figure 2: Plot of Variable Importance using Boruta Algorithm

### 4.3.2 Feature Selection with *fscaret* Library

In order to compare and contrast, two algorithm were performed using *fscaret* library such as Gradient Boosting and pcaNNet. For both the models, the control for the train function was specified as 10 Fold Cross validation with several iterations and the summary function that performs across the sample was a two class summary function.

- Gradient Boosting Model : The choice of this model is due to their reliability and speed in which output is generated. First, the vectors were created using Interaction depth, number of trees to grow, shrinkage parameter and minimum node that stops the tree. Then, it was applied to model using train() function and the output of top 10 variables were obtained as in figure 3.

- Neural Net with Principal Component Analysis - This model works in such a way that the calculated principal components are further used as an input weight for the Neural Network. These weights are calculated using an optimal subset that performs Principal Component analysis. The Importance of the variables are determined and the plot is shown in figure 4.

### 4.3.3 Evaluation of Feature Selection Algorithm

The below table represents the accuracy, sensitivity, specificity, kappa value of the two models with 95% confidence interval. These values are derived from the confusionMatrix() function in R studio.

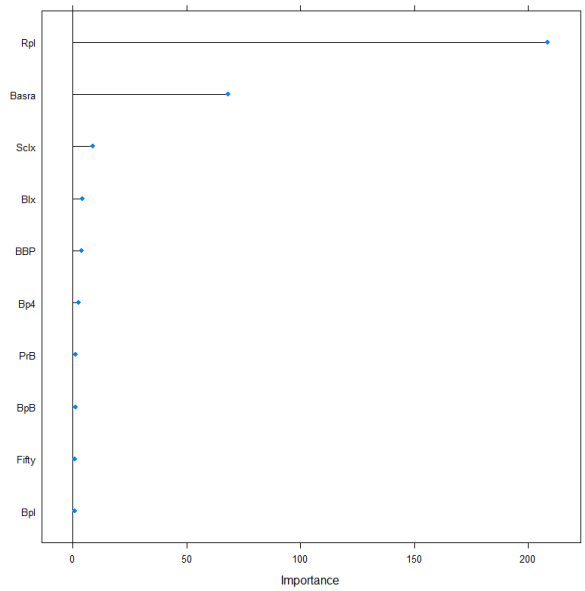


Figure 3: Plot of Variable Importance using GBM

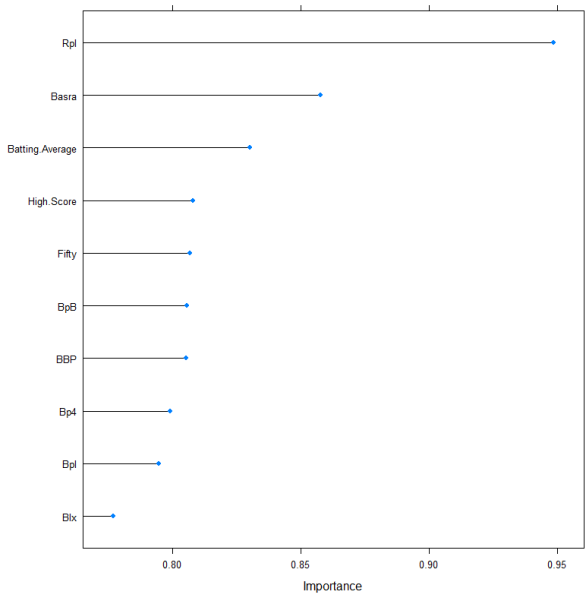


Figure 4: Plot of Variable Importance using pcaNNet

| Model Name | Accuracy | Sensitivity | Specificity | Kappa Value |
|---|---|---|---|---|
| GBM | 0.9686 | 0.9490 | 0.9892 | 0.9372 |
| pcaNNet | 0.9738 | 0.9694 | 0.9785 | 0.9486 |
| Boruta Algorithm | NA | NA | NA | NA |

The most repeated top variables have been made as a subset and considered for further analysis. The below table represents top 10 variables.

| Boruta Algorithm | GBM | pcaNNet |
|---|---|---|
| **Runs Per Innings** | **Run Per Innings** | **Run Per Innings** |
| **BASRA** | **BASRA** | **BASRA** |
| **Ball Per Innings** | Scoring Index | Batting Average |
| Batting Average | **Boundary Index** | High Score |
| Scoring Index | **Boundary Ball Percentage** | **Fifty** |
| **Ball per Boundary** | Ball per Four | **Ball Per Boundary** |
| **Boundary Ball Percentage** | Percent of Boundary | **Boundary Ball Percentage** |
| **Boundary Index** | **Ball per Boundary** | Ball per Four |
| **Fifty** | **Fifty** | **Ball Per Innings** |
| High Score | **Ball Per Innings** | **Boundary Index** |

Figure 5: Top 10 Common Variables from Feature Selection Models

As a conclusion of this section, the author has selected commonly occurred top 10 variables from all the models. The variables are *Runs per Innings, BASRA, Ball Per Innings, Ball Per Boundary, Boundary Ball Percentage, Boundary Index, Fifty.* In addition, *Batting average, Ball Per Four, High Score and Scoring Index* is considered as well since they have repeated two times. In total, 11 variables were used for further analysis.

## 4.4 Modelling - Predicting Contribution Percentage of a Batsman

### 4.4.1 Decision Tree - Conditional Inference Tree

In this project, the ctree() function was executed considering *Effectiveness* as dependent variable and *Runs per Innings, BASRA, Ball Per Innings, Ball Per Boundary, Boundary Ball Percentage, Boundary Index, Fifty, Batting average, Ball Per Four, High Score and Scoring Index* as predictor variables. ctree() has been trained with some of the controls such as, type of test to compute the output class is set as Test Statistics, minimum criterion for split as 0.95 and minimum sum of weight as 20L. In this model, the minimum sum of weight is also a deciding factor on how much tree should be formed. Before all these analysis, the dataset was split into 70 % as training and 30 % as testing dataset, then it was implemented into the train model with the above mentioned parameters. The accuracy was 0.9412 with a kappa value of 0.8824 and p-value as less than 0.05. Figure 6 is the output obtained from the decision tree.

The set of rules derived from the above decision tree is explained below in figure 7.

Based on the set of rules given in below table, a manager can come to the selection panel in such a way that, the player with BASRA greater than 161 and Run Per innings is less than 18 or 21 should not be considered since batsmen with these criterion have always shown negative contribution. Rather, the manager can only focus on player with

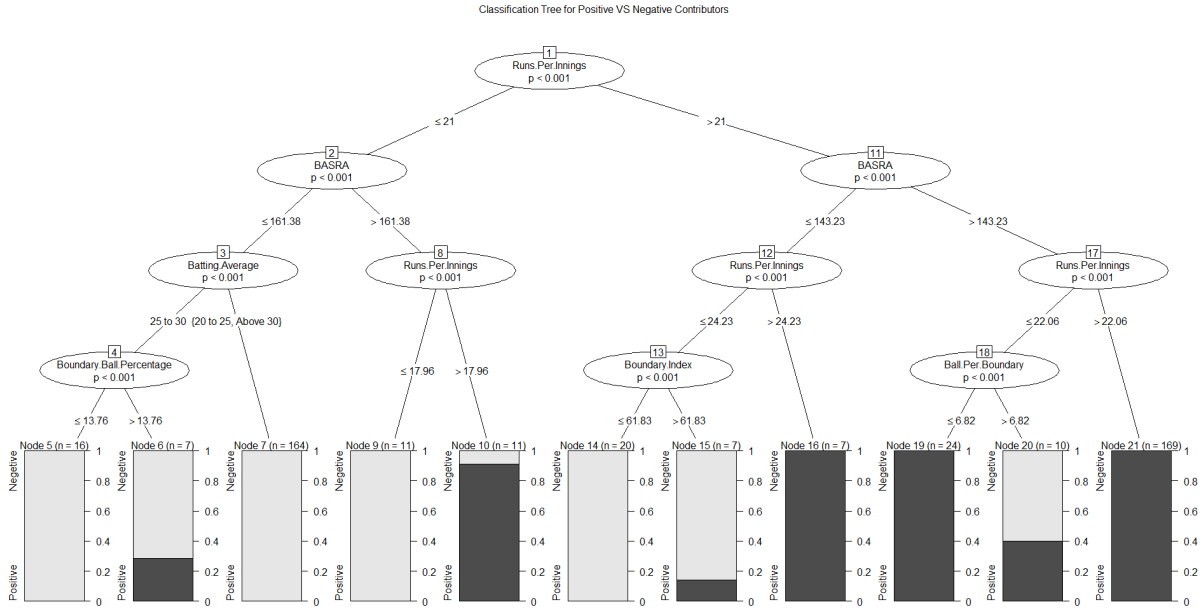


Figure 6: Decision Tree Output

| S.No | Nodes | End Node | If | Then |
|---|---|---|---|---|
| 1. | 1,2,3,4 | 5 | Batsman scored below 21 AND BASRA is less than 161 AND Batting Average is 25 to 30 AND Boundary Ball Percentage is less than 14 | The batsman has shown 100 % negative contribution throughout the series |
| 2. | 1,2,3,4 | 6 | Batsman scored below 21 AND BASRA is less than 161 AND Batting Average is 25 to 30 AND Boundary Ball Percentage is more than 14 | The batsman has shown 30 % positive control throughout the series |
| 3 | 1,2,3 | 7 | Batsman scored below 21 AND BASRA is less than 161 AND Batting Average is below 25 & above 30 | The batsman has shown 100 % positive contribution throughout the series |
| 4 | 1,2,8 | 9 | BASRA is greater than 161 AND Run Per innings is less than 18 or 21 | The batsman has shown 100 % negative contribution throughout the series |
| 5 | 1,2,8 | 10 | BASRA is greater than 161 AND Run Per innings is in between 18 to 21 | The batsman has shown 90 % positive control throughout the series |
| 6 | 1,11,12,13 | 14 | Runs Per Innings is between 21 to 24 with BASRA value as less than 143 AND Boundary Index is less than 62 | The batsman has shown 100 % negative contribution throughout the series |
| 7 | 1,11,12,13 | 15 | Runs Per Innings is between 21 to 24 with BASRA value as less than 143 AND Boundary Index is greater than 62 | The batsman has shown 10 % positive contribution throughout the series |
| 8 | 1,11,12 | 16 | Runs Per Innings is greater than 21 or 24 with BASRA value as less than 143 | The batsman has shown 100 % positive contribution throughout the series |
| 9 | 1,11,17,18 | 19 | Runs Per Innings is 21 or 22 AND BASRA is greater than 143 AND Ball Per Boundary is less than 7 | The batsman has shown 100 % positive contribution throughout the series |
| 10 | 1,11,17,18 | 20 | Runs Per Innings is 21 or 22 AND BASRA is greater than 143 AND Ball Per Boundary is greater than 7 | The batsman has shown 60 % Negative contribution throughout the series |
| 11 | 1,11,17 | 21 | Runs per Innings is greater than 21 or 22 AND BASRA is greater than 143 | The batsman has shown 100 % Positive contribution throughout the series |

Figure 7: Set of Rules derived from Decision tree output

Batsman scored below 21 with Batting Average is below 25 or above 30 and BASRA is less than 161. The next section consist of Naive Bayes algorithm and their model implementation.

### 4.4.2 Naive Bayes Model

The Naive Bayes algorithm was performed using caret and klaR package in R studio. The dataset 1 was split into 70% for training the model and 30% for testing the model quality of the model. The train control parameter used was 10-Fold cross validation with 10 iterations in order to avoid over fitting problems. Then, the training model was created using *Effectiveness* as dependent variable and *Runs per Innings, BASRA, Ball Per Innings, Ball Per Boundary, Boundary Ball Percentage, Boundary Index, Fifty, Batting average, Ball Per Four, High Score and Scoring Index* as predictor variables along with the train control parameter. The below diagram is the output of conditional probabilities derived from Naive Bayes algorithm,

```
Conditional probabilities:
          RpI
Y              [,1]     [,2]
  Negetive  18.35000 2.968891
  Positive  24.83353 3.181115

          Basra
Y              [,1]     [,2]
  Negetive  141.6132 10.62468
  Positive  158.3838 11.34809

          Batting.Average
Y            20 to 25   25 to 30   Above 30
  Negetive  0.77489177 0.19913420 0.02597403
  Positive  0.18099548 0.46606335 0.35294118

          Fifty
Y              [,1]     [,2]
  Negetive   4.099567 3.647936
  Positive  11.755656 9.499475

          BpB
Y              [,1]     [,2]
  Negetive  7.696710 1.374734
  Positive  6.131991 1.028402

          Bp4
Y              [,1]     [,2]
  Negetive  10.586277 2.155549
  Positive   8.277828 1.755144

          BpI
Y              [,1]     [,2]
  Negetive  15.72740 3.249609
  Positive  19.40579 3.168740

          BIx
Y              [,1]     [,2]
  Negetive  60.73939 11.59367
  Positive  75.38941 13.44758

          ScIx
Y              [,1]     [,2]
  Negetive  39.99394 8.115591
  Positive  48.22475 8.451872

          BBP
Y              [,1]     [,2]
  Negetive  13.39584 2.322861
  Positive  16.76303 2.804076
```

Figure 8: Conditional Probability - Naive Bayes

All the variable were continuous except Batting Average variable. Hence, 9 Gaussian (Normal) Distribution was generated by the algorithm in RpI (Runs Per Innings), Basra, Fifty, BpB (Ball Per Boundary), Bp4 (Ball per 4), BpI (Ball per Innings),BIx (Boundary Index) and ScIx (Scoring Index). The [,1] and [,2] in all the mentioned variables are derived based on mean value ([,1] ) and standard deviation([,2]) respectively. Whereas, the Batting Average was calculated using conditional probability formula. In the next section, the model evaluation is presented.

# 5 Model Evaluation

The most important consideration for a model performance is the selection of parameters that are used to train the Model. This subsection is a discussion about the evaluation of the best model that fits the dataset 1. The prediction accuracy of the target variable Effectiveness was 94.12% from decision tree model and 87.23% from Nave bayes model. Apart from accuracy, the other evaluation metric that has to be considered is the Kappa Value. It shows how good a model is fit for the data and the kappa value was 0.8824 and 0.7441 for decision tree model and Naive Bayes model respectively. It is an indication

that both the model has shown good fit over the data. The below figure is a graphical representation of model comparison,

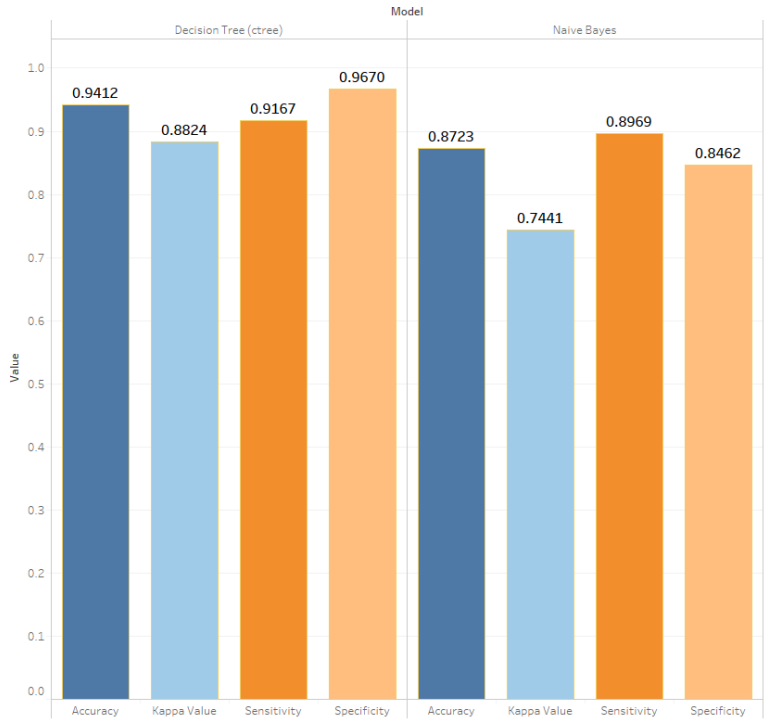


Figure 9: Model Comparison

As a whole, decision tree has performed better than Naive Bayes model. Apart from these two algorithms, there are several other classification method that might bring some good result. There is also a chance that techniques such as Neural Network, Deep learning can gain some more accuracy for the predictive model.

## 5.1 Case Study 1

Q1 : **Does a highly averaged batsman who is capable of scoring more than 20 runs per innings necessarily be a positive contributor of a team?**

It is evident from the decision tree model that there are several other factors that has to be considered in order to estimate a batsmen as a positive contributor. For example, the nodes from decision tree such as 1,2,3,7 states that a *moderately (Below 25) or highly averaged (Above 30) batsmen can also be positive contributor for a team even if he scores less than 20 runs per innings with the fact that the BASRA is less than 161.* In order to make it clear, the author has taken the subset of well known players who scores greater than 20 runs per innings with a high batting average and is presented, evaluated using graphical representation.

For evaluating the below graph in figure 10, player like M.S.Dhoni and Abhinav Mukund can be taken as an example for this case study. Runs per innings is higher for Abhinav Mukund than M.S.Dhoni but Abhinav Mukund has a negative contribution for whichever team he has played in his career whereas M.S.Dhoni has a positive contribution from his side. It is clear that high average and runs per innings are not alone a factor for deciding whether a batsman is a positive contributor or a negative contributor

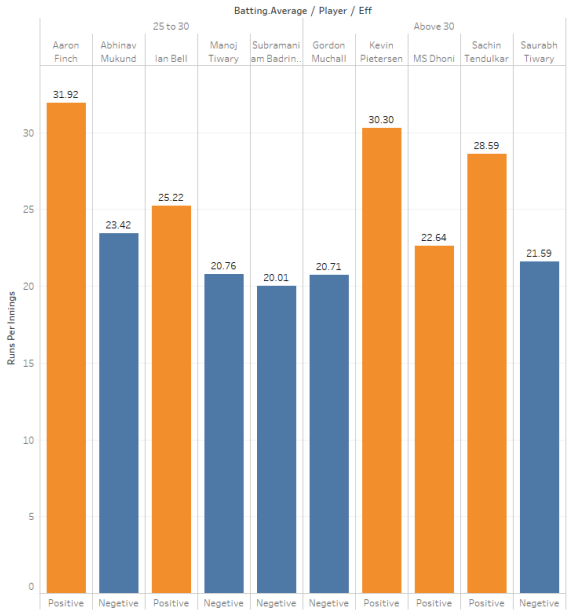


Figure 10: Positive and Negative Contributor classification based on High Average

of a team. Similarly several other players can also be compared and importance to positively contributing player can be given in order to make the game of cricket grow in a reliable way.

## 5.2 Case Study 2

Q2 : **Is the price value of a batsman justified depending on his previous record or consistency of the game?**

In this case study, dataset 2 which consist of Auction price paid (in million) for a player is used. It is merged with dataset 1 and presented for two possible results. The subset of positive contributor is taken using node 1,11,17,21 of decision tree and the subset of negative contributor is taken from node 1,2,8,9 from decision tree. It is then represented in a graph for interpretation.

- Positive Contributor versus price paid : Well known batsmen with positive contribution are compared with the price paid to them during an auction and they are graphical represented below,
- Negative contributor versus price paid : Players with negative contribution is compared and graphically presented below,

From the result of below two graph, firstly the author tries to conclude that players like Chris Woakes, David Weise and Kedar Jadhav has shown negative contribution inspite of being paid more. This is a huge loss to the managers and also affects the team performance. Secondly, from the first graph the author notices that players like Kumar Sangakkara, Owais Shah, Andrew Mcdonald, Mitchell Marsh are paid less even though they have shown positive contribution. So, the author concludes that Chris Woakes, David Weise, Kedar Jadhav can be replaced with positively contributing players like Kumar Sangakkara, Owais Shah, Andrew Mcdonald, Mitchell Marsh in order to form an efficient team that might pay back the money invested on the players.

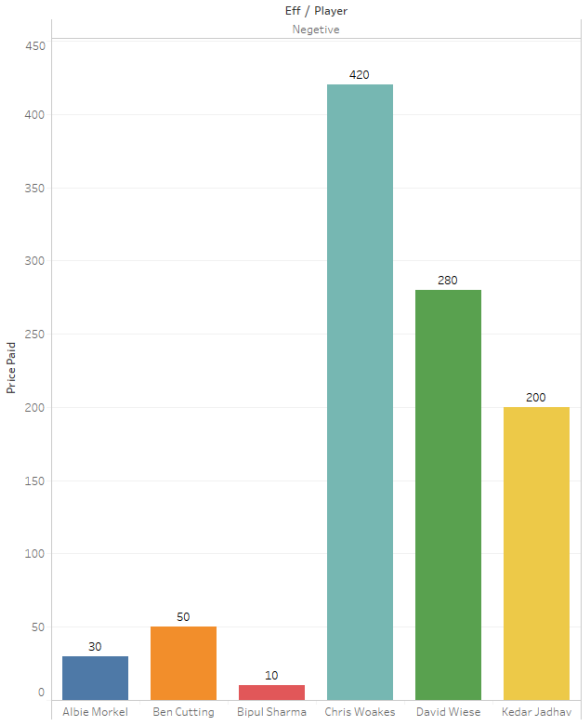


Figure 11: Positive Contributors VS Price

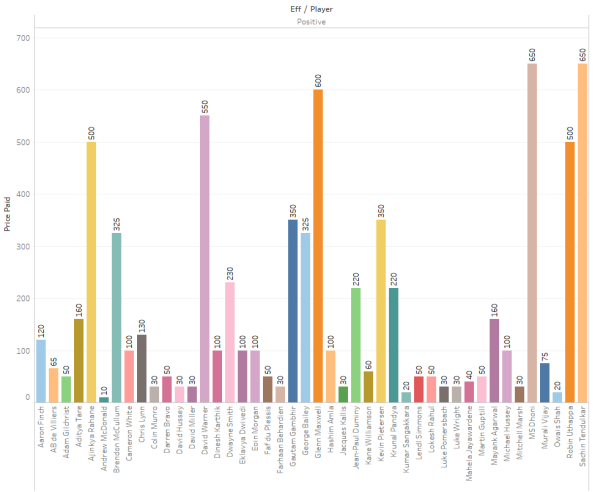


Figure 12: Negative Contributors VS Price

# 6 Conclusion and Future Work

In most of the research papers, the performance of a player was generally determined by batting average, strike rate and number of boundaries hit. In future, it would be better to consider some other factors like boundary ball percentage, BASRA and runs per innings as used in this research project. Also ,the use of data mining and machine learning has to be promoted in the field of cricket for predicting the player performance and their contribution towards a team. In this research concern, player performance dataset was obtained by web scrapping, followed by which feature selection was performed. The features that are highly associated with the dependent variables (Effectiveness) were obtained and classification models were implemented based on the top features derived from the feature selection. The top features were Runs per Innings, BASRA, Ball Per Innings, Ball Per Boundary, Boundary Ball Percentage, Boundary Index, Fifty, Batting average, Ball Per Four, High Score and Scoring Index. The machine learning techniques like decision tree and Naive Bayes were used in this research. Out of these two models, decision tree has shown some promising accuracy and better fit over the data compared to Naive Bayes classification algorithm. Additionally, two case study was presented with the support of decision tree obtained in the modelling stage. The result of case study 1 proves that a batsmen with high average and a score of above 20 runs per innings alone does not make him as a positive contributor, rather there are several other factors such as BASRA, Boundary Ball percentage and Boundary Index has to be considered while calculating a player contribution towards a team. Then, the case study 2 reveals about the price paid for a positive and negative contributor during an auction. This was a small attempt to help the managers deciding the best players for a positive team.

Some of the techniques like artificial neural network, support vector machines, deep learning and ensemble methods like bagging, boosting, random forest can be implemented in future for better accuracy and prediction result. Feature selection using SVM, Random forest can also be implemented in future. Apart from the modelling, the main drawback of this research is that it has not included the bowling and fielding performances. Hence, similar research can be carried out on bowler dataset to define how well a bowler can perform in a game regardless of the external conditions. Also, external factors like weather, kind of pitch the batsmen is playing, crowd attendance and lighting condition are not considered in this research. In future, a prediction model can be developed for influence of external conditions on player performance and contribution.

# Acknowledgements

I would like to thank Dr. Eugene O'Loughlin who helped to come up with my thesis work. He gave me various suggestions on my project. Then, I would like to thank my Father and Mother for giving me this oppurtunity.

# References

Ahmed, F., Deb, K. and Jindal, A. (2013). Multi-objective optimization and decision making approaches to cricket team selection, *Applied Soft Computing* **13**(1): 402–414.

Barr, G. and Kantor, B. (2004). A criterion for comparing and selecting batsmen in limited overs cricket, *Journal of the Operational Research Society* **55**(12): 1266–1274.

Beaudoin, D. and Swartz, T. B. (2003). The best batsmen and bowlers in one-day cricket, *South African Statistical Journal* **37**(2): 203.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Dyte, D. and Clarke, S. R. (2000). A ratings based poisson model for world cup soccer simulation, *Journal of the Operational Research society* **51**(8): 993–998.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of machine learning research* **3**(Mar): 1157–1182.

Hothorn, T., Hornik, K. and Zeileis, A. (2015). ctree: Conditional inference trees, *cran. r_project* .

Iyer, S. R. and Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection, *Expert Systems with Applications* **36**(3): 5510–5522.

Kansal, P., Kumar, P., Arya, H. and Methaila, A. (2014). Player valuation in indian premier league auction using data mining technique, *Contemporary Computing and Informatics (IC3I), 2014 International Conference on*, IEEE, pp. 197–203.

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J. and Shmulevich, I. (2009). Fewer permutations, more accurate p-values, *Bioinformatics* **25**(12): i161–i168.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection, *Artificial intelligence* **97**(1-2): 273–324.

Lemmer, H. H. (2008). An analysis of players' performances in the first cricket twenty20 world cup series, *South African Journal for Research in Sport, Physical Education and Recreation* **30**(2): 71–77.

Loeffelholz, B., Bednar, E., Bauer, K. W. et al. (2009). Predicting nba games using neural networks, *Journal of Quantitative Analysis in Sports* **5**(1): 1–15.

Ma, B. L. W. H. Y. and Liu, B. (1998). Integrating classification and association rule mining, *Proceedings of the 4th*.

Pawlak, Z. (2002). Rough sets, decision algorithms and bayes' theorem, *European Journal of Operational Research* **136**(1): 181–189.

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.

Scarf, P., Shi, X. and Akhtar, S. (2011). On the distribution of runs scored and batting strategy in test cricket, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**(2): 471–497.

Suzuki, A. K., Salasar, L. E. B., Leite, J. and Louzada-Neto, F. (2010). A bayesian approach for predicting match outcomes: the 2006 (association) football world cup, *Journal of the Operational Research Society* **61**(10): 1530–1539.

Wickramasinghe, I. P. (2014). Predicting the performance of batsmen in test cricket.

Zhang, Z. (2016). Naïve bayes classification in r, *Annals of Translational Medicine* **4**(12).

Zimmermann, A., Moorthy, S. and Shi, Z. (2013). Predicting ncaab match outcomes using ml techniques–some results and lessons learned, *Machine Learning and Data Mining for Sports Analytics Workshop (MLSA-13). Prague, Czech Republic*, Vol. 27.

# A Appendix 1

| Variable | Description | Data Type | Missing value | Mean | Std.Deviation |
|---|---|---|---|---|---|
| Player | Name of the player | String | 0 | 808.50 | 466.54 |
| Match | Total Match represented | Numerical | 0 | 60.66 | 45.81 |
| Innings | Total match got chance to bat. | Numerical | 0 | 44.58 | 42.53 |
| Not Outs | No. of Matches not dismissed out | Numerical | 1001 | 27.69 | 19.10 |
| Runs | Total Runs | Numerical | 0 | 801.12 | 1087.12 |
| High Score | Best score in single match | Numerical | | 54.12 | 32.62 |
| Batting Average | (Total Runs) / (No.of matches dismissed) | Numerical | 195 | 568.28 | 336.59 |
| Balls Played | Balls faced during an innings | Numerical | 0 | 653.63 | 831.33 |
| Strike Rate | ((Total Runs) / ( Balls Played)) * 100 | Numerical | 0 | 112.19 | 23.46 |
| 100's | A Century (Rare in Limited Over Cricket) | Numerical | 3718 | 1.19 | 0.72 |
| 50's | Greater than 50 & Less than 100 runs. | Numerical | 2772 | 11.02 | 14.58 |
| 30's | Greater than 30 & Less than 50 runs. | Numerical | 2048 | 16.77 | 17.64 |
| 0's | Duck Outs | Numerical | 1608 | 11.68 | 7.54 |
| Four | Ball travelled to boundary rope along the ground | Numerical | 612 | 149.55 | 96.02 |
| Six | Ball travelled in the air crossing the boundary rope | Numerical | 1314 | 66.48 | 48.17 |
| Balls Per boundary (BPB) | Balls faced by a batsman to hit each boundary | Numerical | 517 | 377.99 | 197.70 |
| Balls per Four (Bp4) | Balls faced by a batsman to hit each four | Numerical | 612 | 394.12 | 262.34 |
| Balls per Six (Bp6) | Balls faced by a batsman to hit each Six | Numerical | 1314 | 504.32 | 343.81 |
| Percent of Boundary | Total percent of four and six | Numerical | 517 | 551.74 | 333.69 |
| Percent of Four | Total Percent of Four | Numerical | 612 | 504.13 | 314.04 |
| Percent of Six | Total Percent of Six | Numerical | 1314 | 449.04 | 319.27 |
| Boundary Ball Percentage (BBP) | Probability that a batsman hit a boundary on particular ball | Numerical | 517 | 422.86 | 274.25 |
| Runs Per Boundary (RpB) | Runs scored in between hitting a 4 or 6. | Numerical | 517 | 52.24 | 30.28 |
| Boundary Index (Bix) | Runs Per Boundary * Boundary Ball Percentage | Numerical | 517 | 600.20 | 370.69 |
| Runs Per Non-Boundary | Runs scored in between each run (1,2,3) | Numerical | 12 | 46.95 | 10.28 |
| Scoring Index (Scix) | Calculated field using Batsman score, Average and Strike Rate | Numerical | 529 | 648.99 | 384.47 |
| Runs Per Innings (Rpi) | (Total Runs) / (Total Innings) | Numerical | 0 | 13.79 | 7.93 |
| Ball Per Innings (Bpi) | How many balls does a particular batsman is capable of facing ? | Numerical | 0 | 11.83 | 6.21 |
| Batting Position | Whether a batsman is a top order or a middle order | Numerical | 0 | 5.92 | 2.76 |
| BASRA | Aggregate of Batting Average and Strike Rate | Numerical | 0 | 716.39 | 416.23 |
| Effectiveness (Eff -Target Variable) | Calculated using pervious performance based on their contribution to the team | Numerical | 0 | -37.73 | 39.72 |

Figure 13: Dataset 1 Total row count of raw data- 3899 (History of all players who player limited over cricket)

# B Appendix 1

| Variable | Description | Data Type |
|---|---|---|
| Player | Name of the Player | String |
| Type | Bowler/Batsmen/All Rounder | String |
| Price Paid | Base price of a player during an auction | Numeric |
| Origin | Foreign or Indian | String |
| Who Bought the Player | Team which the player currently playing | String |

Figure 14: Dataset 2  Past 4 year IPL auction dataset (Indian Premier League)