RESEARCH

includes research articles that focus on the analysis and resolution of managerial and academic issues based on analytical and empirical or case research

On Classification of All-rounders of the Indian Premier League (IPL): A Bayesian Approach

Hemanta Saikia and Dibyojyoti Bhattacharjee

Executive Summary

An all-rounder can take an imperative role in any version of the game of cricket, whether it is a test match or any other limited-over format of the game. The study classifies the performance of all-rounders who participated in IPL based on their strike rate and economy rate. Based on the factors mentioned, the all-rounders can be divided into four non-overlapping classes, viz., *Performer*, *Batting All-rounder*, *Bowling All-rounder*, and *Under-performer*. Several predictor variables that are supposed to influence the performance of all-rounders are considered. Step-wise multinomial logistic regression (SMLR) is used to identify the significant predictors. Samples of six incumbent all-rounders who had not participated in the first three seasons of IPL are considered. The significant predictors were then used to predict the expected class of an incumbent all-rounder using naïve Bayesian classification model. The relevant data were collected from the websites, *www.cricinfo.org* and *www.cricketnirvana.com*.

The key points of this study are as follows:

- The training sample is populated with 35 all-rounders who had performed in the first three seasons of IPL.
- Two variables, viz., strike rate (number of runs scored per 100 balls faced) and economy rate (average number of runs scored per over against the bowler) are used to classify the all-rounders as follows:
  - *i)* *Performer*: An all-rounder with strike rate above median and economy rate below median.
  - *ii)* *Batting All-rounder*: An all-rounder with strike rate above median and economy rate above median.
  - *iii)* *Bowling All-rounder*: An all-rounder with strike rate below median and economy rate below median.
  - *iv)* *Under-performer*: An all-rounder with strike rate below median and economy rate above median.
- The step-wise multinomial logistic regression (SMLR) was used to identify the significant variables that are actually responsible for classification of the all-rounders.
- The strike rate in ODI, strike rate in Twenty-20, economy rate in ODI, economy rate in Twenty-20 and bowling type (Spin or Fast) of the all-rounders are found to be significant in determining the class of an all-rounder.
- The naïve Bayesian classification model is used for forecasting the expected class of all-rounders based on the significant predictors for six incumbent all-rounders who had played only in fourth season of IPL.
- The prediction done before IPL IV was then compared with the actual situation at the end of the tournament. It is found that four predictions were performed correctly out of the six.

This model would be useful for the participating teams' management while deciding the bid of an all-rounder in the upcoming season of IPL as per their requirement.

KEY WORDS

Naïve Bayesian Classification

Multinomial Logistic Regression

Cricket

Performance Measurement

Data Mining

Decision Making

Cricket in India has captured the attention of most of the sports enthusiasts and the media that cover its developments on a regular basis. It is managed reasonably well by the Board of Control for Cricket in India (BCCI) compared to the governing bodies of other sports which lack adequate institutional support and planning (Ramani, 2008). When India won the inaugural Twenty-20 World Cup in 2007, Indians had several reasons to be happy about. Firstly, it was her major success after the Cricket World Cup victory in 1983. Secondly, in the same year, in the 50-over version of World Cup cricket, India had performed poorly, getting eliminated in the first round itself. The deep wounds of the fans, fresh in their memories, were more than just healed when M S Dhoni and his men clinched the inaugural Twenty-20 World Cup title.

In April 2008, the game of cricket got a new dimension when BCCI initiated the Indian Premier League (IPL), a Twenty-20 cricket tournament to be played among eight domestic teams, named after eight Indian states or cities but owned by franchisees. The franchisees formed their teams by competitive bidding from a collection of Indian and international players and the best of Indian up-coming talent. Team owners bid for the services of cricketers for a total of US $42 million. Each team could purchase a maximum of eight overseas players; though only four would be considered in the playing eleven (Rastogi and Deodhar, 2009). The franchisees bid for the salaries that they were ready to offer to the players. Each player had a base price fixed by the IPL authorities and there was no upper limit. However, the salary offer was valid for three years only. From the fourth season, two new teams joined IPL. The league expanded in terms of the number of teams and the number of matches played. The salary offers to the available cricketers were also renewed through fresh bidding. But such change should be related to the performance of cricketers in the previous seasons of IPL and in other domestic and international tournaments.

Of the different dexterities required to become a cricketer, batting and bowling are undoubtedly the prime skills in the game of cricket. Though a balanced cricket team has specialist batsmen and bowlers, yet players with reasonably good performance with both the bat and the ball are always vital assets to the teams. Such players are called all-rounders. Bailey (1989) defined an all-rounder as a player who is able to command a place in his side for either his batting or his bowling. This definition actually followed a previous argument by Cardus (1978), who stated that the test for a great all-rounder is, 'Would he be picked to play in a test match only for his batting or for his bowling?' Sometimes, you find players who are capable of both batting and bowling to a high standard. "These players are called all-rounders and are worth their weight in gold as far as selectors are concerned. One good example is England's Andrew Flintoff, a player good enough to be selected both as a batsman and as a bowler but only occupying one place in the team" (Knight, 2006). The all-round performance of those cricketers who can bat and bowl regularly have played a significant role because the captain of a team would utilize those players solely as a batsman or a bowler whenever he liked during a match (Van Staden, 2008). Such players are often referred to as utility players as they have a chance to perform with the ball in case they fail to bring laurels to their team with their willow and *vice versa*. However, according to Rundell (2006), the criteria mentioned by Cardus (1978) are rigorous; in practice, many of the great all-round performers of the past are remembered primarily for their skill in one or the other department of the game. Essentially, an all-rounder is better at bowling than batting or *vice-versa*. Very few are equally good at both and hardly any have been outstanding at both. One of the main constraints to becoming a recognized all-rounder is that batsmen and bowlers "peak" at different ages. Batsmen tend to reach their peak in their late twenties after their technique has matured through experience. Conversely, fast bowlers often reach their peak in between early to mid-twenties when they are at the height of their physical prowess. Other bowlers, mostly spinners, even fast bowlers who can "swing" the ball, are most effective in the later part of their career. Thus, the terms "bowling all-rounder" and "batting all-rounder" have come into use (wikipedia. com).[1] A genuine all-rounder is an asset to the team, but in the absence of such all-rounders in modern cricket, the responsibility of all-rounders are shared by batting all-rounder(s) or/and bowling all-rounder(s).

The paper evaluates the performance of the players who both batted and bowled in the first three versions of IPL and classified the players based on their performance. The players are classified into four classes of all-round-

[1] Wikipedia, the free encyclopedia (*www.wikipedia.com*) accessed on 16th October, 2010.

ers — genuine all-rounder (performers both with their bat and ball), bowling all-rounder, batting all-rounder, and under-performer. The factors responsible for the classification are identified using step-wise multinomial logistic regression. Later, a Bayesian classification method is developed from the training sample of the all-rounders who participated in IPL I, IPL II, and IPL III. The Bayesian classification criteria can then be utilized to predict the class in which an incumbent all-rounder is expected to fall. The external validity of the model is tested by predicting the performance of the incumbent all-rounders, who participated in IPL IV. The predicted performance is then compared with the actual performance of those incumbent all-rounders obtained at the end of IPL IV. Such a model can be helpful to franchisees in deciding about which all-rounder to bid for, depending on the requirement of their team.

REVIEW OF LITERATURE AND OBJECTIVES OF THE STUDY

Twenty-20 is the latest version of cricket. IPL has emerged as a focus of discussion and a hot-spot for research workers of various disciplines, viz., Economics, Management, Decision Science, Finance, Human Resource and so on. This is mainly because the players' salaries are decided through auction, which has led to a rare opportunity of obtaining the values of the players in monetary terms. Consequently, considerable amount of literature on Twenty-20 cricket has evolved considering its age and compared to the literature available on the other versions of cricket.

Vig (2008) studied the implications of having two cricket leagues in India, viz., Indian Cricket League (ICL) and Indian Premier League (IPL). Ramani (2008) reported IPL as a "…distorted form of commodity and consumer excess." An economic model was developed by Das (2008) to capture the inherent conflict between cricket boards, players, and international club line game sponsors like ICL or IPL. Lenten, Geerling and Konya (2009) compared a range of cross-sectional models to study the factors determining the performance of cricketers in different forms of the game including IPL.

Performance measurement and classification of players based on their performance can be considered as a researchers' delight irrespective of the sports under consideration. Bloomfield, Polman and O'Donoghue (2007) have done it for English Football Association Premier League soccer; Carvajal *et al* (2009) classified elite Cuban baseball players into five categories: Clerke (1997) for soccer and cricket; Gabbett (2002) for rugby; McGee and Burkett (2003) for football; Schwandt, Glenn and Triantis (2007) for atheletics, etc.

In cricket, however, relevant literature pertaining to performance-based classification of players is not that rich. Traditionally, different measures like batting average and strike rate are mostly used to understand the performance of batsman in cricket. On the other hand, bowling average, economy rate, and bowlers' strike rate are used to measure the performance of bowlers. Lewis (2005) stated that the existing measures for players' performance in cricket are unable to access the true ability of the players. Accordingly, an alternative performance measure was proposed, extending the Duckworth/Lewis method that can even take into consideration the situation in which a player performs. Lemmer (2004; 2006) developed a classification scheme for batsmen and bowlers using performance data of one-day international (ODI) matches and Test cricket. Soon Lemmer (2008) discussed the performance of cricketers in the first Twenty-20 World Cup. In terms of the all-round performance of those players who bat and bowl regularly in the first edition of IPL, Van Staden (2008) used the term ideal all-rounder, batting all-rounder, and bowling all-rounder to characterize the performance of all-rounders. Van Staden (2009) developed a performance measure for cricketers in Twenty-20 cricket considering data from IPL I. Tan and Ramachandran (2010) ranked the all-rounders in test cricket, both past and present, in accordance with a mathematical formulae derived from batting and bowling records. Brettenny (2010) reviewed the different existing measures of players' performance in cricket and used them to select players for a fantasy league using the binary integer programming model. Some bidding strategies in IPL are also discussed in Singh, Gupta and Gupta (2011). The franchisees of the different teams in IPL are actually locked in an optimization problem. On the one hand, they are to form the strongest possible team, while on the other hand, they are constrained by the resources, i.e., cost, the total of which cannot exceed US $42 million and also not more than eight foreign players can be purchased. Existence of all-rounders in a team can contribute significantly towards the team's success. But as perfect all-rounders are few in number, they are ought to be highly priced. Since the budget is limited, a franchisee could afford perfect

all-rounders only at the cost of other expensive players. The bowling all-rounders (mainly bowlers who are also known for their batting talent) and the batting all-rounders (specialist batsmen who can occasionally bowl successfully) can be other options to the franchisee in the absence of perfect all-rounders. Thus, it is important that the team franchisees decide precisely while selecting all-rounders for their team. The concern of this paper is to classify the all-rounders who participated in IPL, based on their performance into any one of the four classes, viz., Performer (perfect all-rounder), Batting all-rounder, Bowling all-rounder, and Under-performer. The factors responsible for such classification shall also be recognized. Once the responsible factors are identified, a classifier can be developed which, with necessary inputs of an incumbent all-rounder, can predict his expected class, before he actually plays the tournament. Such a model can be helpful to the team management to decide which player to bid for depending on their requirement and available fund.

CLASSIFICATION OF PLAYERS IN THE TRAINING SAMPLE

Data

Two variables, viz., strike rate of the batsman, which is the number of runs scored per 100 balls faced and economy rate of the bowler, which is the average number of runs scored per over, are used as performance measures of the all-rounders. First, it may be recognized that the strike rate and economy rate are quantities of opposite nature. The strike rate is a direct quantity — greater the strike rate, more effective the batsman is supposed to be for Twenty-20 format. The economy rate on the other hand, is an inverse quantity — the smaller the economy rate, better is the bowler for the current format of cricket. Ideally, an all-rounder is supposed to have a high strike rate and low economy rate to be effective in the Twenty-20 cricket. Based on the performance of the all-rounders in the first three seasons of IPL, they can be divided into four classes, viz., performer, batting all-rounder, bowling all-rounder, and under performer:

i) An all-rounder with strike rate above median and economy rate below median is a *performer*, which is the ideal situation.
ii) An all-rounder with strike rate above median and economy rate above median is a *batting all-rounder.*
iii) An all-rounder with strike rate below median and economy rate below median is a *bowling all-rounder.*
iv) An all-rounder with strike rate below median and economy rate above median is termed as *under-performer*, which is the worst case.

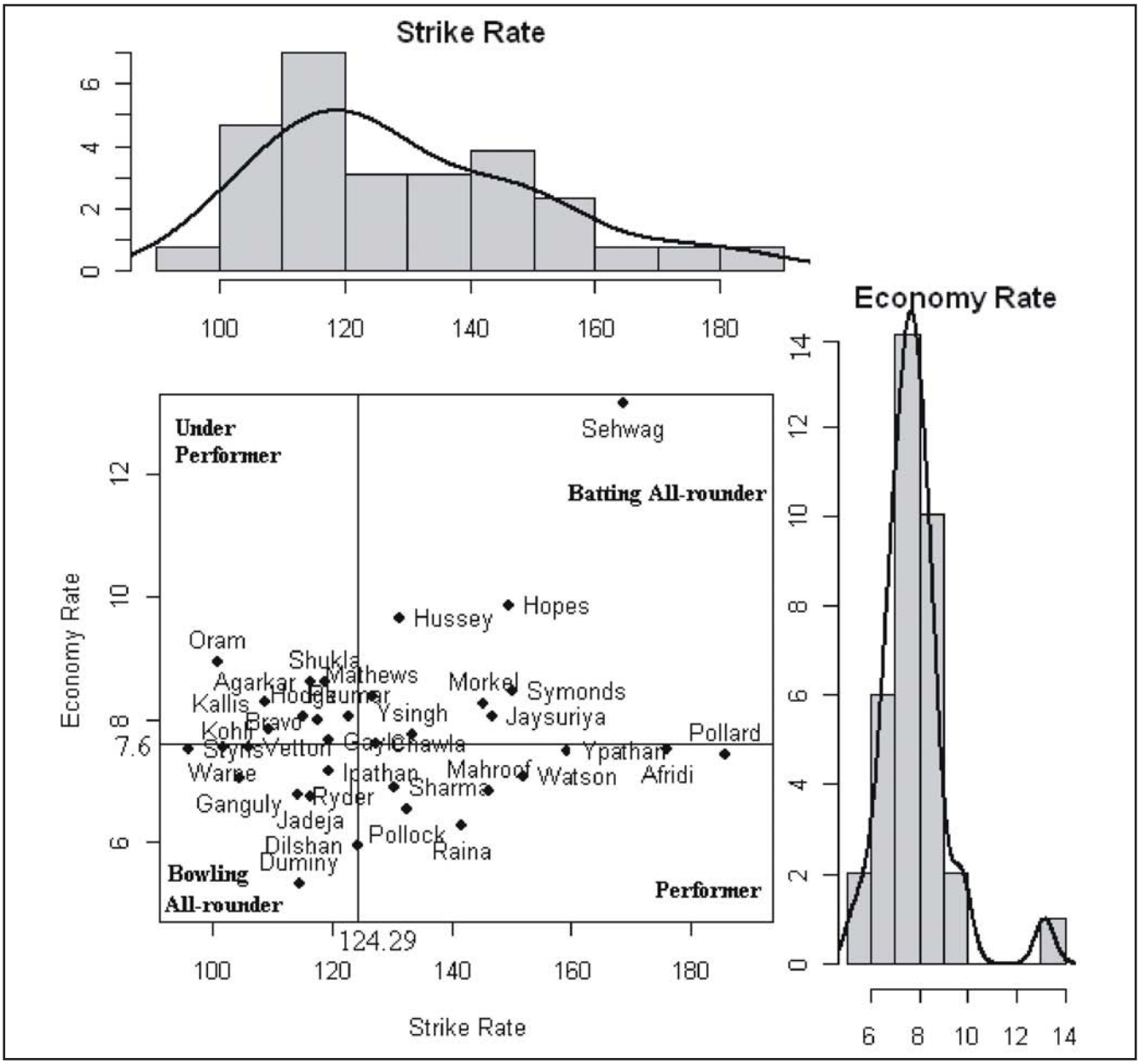
Median is that value of the variate which has equal number of observations both above and below it and is not influenced by extreme values. Median is a better measure of central tendency for non-normal data. Keeping in mind the aforesaid reasons, median is used as the criteria for classification. The dependent variable here, i.e., the type of all-rounder is categorical in nature. Expectations regarding level of performance cannot be gauged fairly from only one match; therefore individual performances across a series of matches are required to provide a suitable frame of reference (Bracewell and Ruggiero, 2009). The nature of the professional sport ensures that the majority of individuals will experience sufficient match-play to enable this type of methodology to be deployed (Bracewell and Ruggiero, 2009). The training sample (Appendix A) considered here is populated with the all-rounders, who had performed in the first three IPLs. However, only those players who satisfied all the following conditions are spaced in the training sample:

- The player was in the playing eleven for at least 5 matches in IPL
- The player has bowled at least 10 overs in IPL
- The player has faced at least 100 balls in IPL.

As discussed earlier, an all-rounder will fall in any one of the categories viz. *performer*, *batting all-rounder*, *bowling all-rounder,* and *under-performer* based on his strike rate and economy rate. The graphical depiction of the classification is given in Figure 1 .

The training sample initially consisted of 40 all-rounders, but in case of some of the all-rounders, the values of some of the independent variables (discussed in the next section) are not available. For example, Y Venugopal Rao has not bowled in any ODI and hence his economy rate in ODI is not available. Manpreet Singh Gony has not batted in any ODI and so his strike rate in ODI was not available; similarly, K Goel has also not played in any ODI and so his information for ODI was not available, at the time when the data was collected. Ultimately, the training sample size came down to 35. The information about the all-rounders for the variables considered earlier was collected from the website (*www.cricinfo.org)* on

Figure 1: Graphical Display Categorizing All-rounders of the Training Sample



Source: The graph is drawn in R with data from Appendix B.

April 4, 2011 just after the completion of the World Cup and before the start of IPL IV.

Variables

Several independent variables that are supposed to influence the said classification are considered. Some of them are nominal (e.g., batting hand either left or right); some are discrete (e.g., IPL team, country, etc.) while others are continuous (e.g., age, strike rate in Twenty-20, etc.) in nature. A brief description of the different independent variables is given in Box 1 and the data can be seen in Appendix A.

Identification of Significant Factors

This study attempts to identify the various factors that are associated with the classification of the cricketers into the four classes of all-rounders as discussed earlier. To determine the variables that are influential in shaping the class to which an all-rounder may belong, step-wise multinomial logistic regression is used. Such regression techniques are used whenever the dependent variable is categorical in nature and is classified into more than

Box 1: Description of Independent Variables

Age: Demographic variable measuring the age of the player. It is the number of years completed at the end of IPL III.

Batting hand: Nominal variable, i.e., either left or right. It is the dominating hand of the player while batting. The binary code '0' represents left hand batsmen and '1' represents right hand batsmen.

Bowling hand: This is a nominal variable (i.e., either left or right). It is the throwing arm of the bowler when he comes to bowl. The binary code '0' represents left hand bowler and '1' represents right hand bowler.

Type of bowler: The type of bowling is characterized into two groups either fast or spin. Medium fast bowlers are considered under the fast bowler category. The code '1' is used for fast bowler and '2' for spin bowler.

ODI matches played: Measures the experience of the player in terms of number of international one day matches in which he was in the playing eleven.

Twenty-20 matches played: Measures the experience of the player in Twenty-20 cricket. It is the number of international Twenty-20 matches in which he was in the playing eleven.

International cricket experience: This is also another measure of experience. It is measured in years which count the international career of the player in terms of years.

Bid amount in IPL: It is the amount of money in dollars for which a given player was auctioned in IPL.

IPL team: The team for which the player played in IPL. There are a few players who changed their team in IPL III, but the training sample does not contain any of them.

Country: The country to which the player belongs.

Average batting position in IPL: IPL matches have seen huge changes in the batting order and hence the average batting positions of a player for all the IPLs taken together are considered.

Strike rate in ODI: Career strike rate of the player in One-Day International matches.

Economy rate in ODI: Career economy rate of the player in One-Day International matches.

Strike rate in Twenty-20: Career strike rate of the player in international Twenty-20 matches.

In case the player has not batted in any Twenty-20 international matches, his strike rate at various other domestic Twenty-20 matches is considered.

Economy rate in Twenty-20: Career economy rate of the player in Twenty-20 international matches. In case the player has not bowled in any Twenty-20 international matches, his economy rate at various other domestic Twenty-20 matches is considered.

two groups, as the case here. They are quite common, especially for demographers, who desire to identify the significant predictors when the response variables are categorical in nature with more than two categories. Some relevant references include Indongo and Naidoo (2009); Njogu and Martin (2006); Dwivedi, Ram and Reshmi (2007); and Fraboni and Rosina (2006). Here step-wise multinomial logistic regression is used to determine the significant predictors. In a step-wise regression, the predictors entered in the model are based on a purely mathematical criterion. An initial model is defined which contains only the constant (b_0). Then search is on for a predictor, out of those available, that best predicts the outcome variable. Generally, the predictor with the maximum value of the Wald statistic[2] is chosen. If inclusion of this predictor significantly improves the ability of the model in terms of predicting the outcome, then the predictor is retained in the model and search for a new predictor starts. The process continues until none of the remaining predictors shows any significant value of the corresponding Wald statistic. At each inclusion, a check is made if any of the predictors have lost its significance because of a new predictor being included in the model (Field, 2009). Due to the existence of intercorrelations between variables, the variance explained by certain variables may change as and when a new variable or variables are entered in the equation. If this takes place, then step-wise method would remove the weakened variables. Though step-wise regression models are criticized at times, thus restricting its use, yet, in case of an exploratory study, where it is important to find a model to fit a data, such a method should be recommended (Agresti and Finlay, 1986; Menard, 1995).

Here, the dependent variable is a type of all-rounder.

[2] Wald statistic is generally used to test if the coefficient for a predictor (say, b) is significantly different from zero. In case it is significantly different from zero then we can assume that the corresponding predictor is making a significant contribution to the prediction of the outcome. The statistic is given by, $W = \frac{b}{SE_b}$ which follows the χ^2 *(chi-square)* distribution.

As discussed, the incumbent all-rounder will fall in any one of the class of all-rounders. The players are classified as: performer (Y=1), batting all-rounder (Y=2), bowling all-rounder (Y=3), and under-performer (Y=4). Also, let X_{i1} be the value of the first independent variable when Y takes the value i, obviously i =1, 2, 3 and 4. Similarly, we can have X_{i2},…, X_{ik} as the other independent variables. Thus, we can have the multinomial logistic regression equation as,

$$Prob(Y = i) = \frac{\exp(b_0 + b_1X_{i1} + b_2X_{i2} + ... + b_kX_{ik})}{1 + \exp(b_0 + b_1X_{i1} + b_2X_{i2} + ... + b_kX_{ik})} \text{ for } i = 1, 2, 3 \text{ and } 4 \quad (1)$$

where, b_0, b_1, b_2,…,b_k are the parameters of the model to be estimated from the data (Appendix A).

Table 1: Information about the Fitted Model

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	Df	Sig.
Intercept Only	96.730			
Final	11.601	76.459	27	0.000

Nagelkerke's R-Square = 0.937 (Software used SPSS version 17)

In case of any regression model, it is essential to have a measure that can be used to identify how well the model fits the data. Hosmer and Lemeshow (1989) defined a measure that could be used to compute the strength of a fitted logistic regression equation given by,

$$R_L^2 = \frac{-2LL(\text{model})}{-2LL(\text{intercept only})} \quad (2)$$

where LL (intercept only) implies log-likelihood of the model before any predictor was entered and LL (model) implies the log-likelihood of the model after entering the predictors. The value of R_L^2 lies between 0 (implying that the predictors are useless in predicting the outcome variable) and 1 (means that the model predicts the outcome variable perfectly) (Field, 2009). However, this measure does not take into consideration the sample size, which is an important factor, while taking such an important decision about the model. Accordingly, Cox and Snell (1989) developed another measure that took into consideration the sample size as well. The measure was defined as,

$$R_{CS}^2 = 1 - e^{-\frac{2}{n}(LL(\text{model}) - LL(\text{intercept only}))} \quad (3)$$

However, this statistic never reaches its theoretical maximum of 1 and so Nagelkerke (1991) suggested the following change to the previous formula.

$$\rho_N^2 = \frac{R_{CS}^2}{1 - e^{\left(\frac{2LL(\text{intercept only})}{n}\right)}} \quad (4)$$

The value of ρ_N^2 may be termed as Nagelkerke's value of Pseudo R-square and can be used to measure the amount of uncertainty that is been explained by the logistic regression model.

Considering data from the training sample, the model is fitted. The information about the fitted model provided in Table 1, shows that the independent variables explain the said classification in a significant manner. The Table indicates how well the model fits the data where a smaller -2loglikelihood value implies that the model fits the data better. In addition to this, the Nagelkerke's value of Pseudo R-square is 0.937 which implies that the data provides a good fit to the multinomial logistic regression equation.

Table 2: Results of Likelihood Ratio Tests to Identify the Significant Variables

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	13.282	0.000	0	0.000
Strike rate in ODI	51.782	36.376	3	0.000
Economy rate in Twenty-20	33.405	9.777	3	0.047
Strike rate in Twenty-20	36.484	11.601	3	0.003
Economy rate in ODI	39.09	15.186	3	0.000
Bowling type	38.5	13.296	3	0.000

(Software used SPSS version 17)

In logistic regression, the observed and predicted values can be used to assess the fit of the model. Therefore, the log-likelihood is based on summing the probabilities associated with the predicted and actual outcomes (Tabachnick and Fidell, 2001). The log-likelihood statistic is an indicator to ensure how much unexplained information is there after the model has been fitted (Field, 2009). Here the likelihood ratio-tests identified those variables using chi-square statistics which are found to be significant in step-wise multinomial logistic regression analysis. It can be computed as,

$$\chi^2_{df} = 2\left[LL(New) - LL(Intercept\,only)\right] \quad (5)$$

where, $df = k_{new} - k_{intercept\ only}$

The *LL (intercept only)* in the model indicates that only the constant is included in logistic regression and *LL (new)* indicates that one or more predictors are added in the step-wise logistic regression step-by-step to improve the model (Field, 2009). For chi-square statistics, the degrees of freedom are calculated by subtracting number of parameters in the baseline model (i.e., intercept only) from the number of parameters in the new model. The number of parameters, *k*, in the baseline model (i.e., intercept only) will always be 1 because the constant is the only parameter to be estimated.

From Table 2, it is seen that the *p*-value corresponding to strike rate in ODI cricket, economy rate in ODI cricket, strike rate in Twenty-20 cricket, economy rate in Twenty-20 cricket, and bowling type are significant in determining the class to which an incumbent all-rounder is supposed to plummet. The remaining variables are found to be insignificant and hence are dropped from further analysis. The calculations are done using SPSS 17.0.

Naïve Bayesian Classification

The Naïve Bayesian classifier is a probabilistic classifier, based on the Bayes theorem. It can predict class membership probabilities such as the probability that a given subject belongs to a particular class. It can handle an arbitrary number of independent variables whether it is continuous or categorical (Zhang, 2004). The naïve Bayes classifier assumes that the effect of a variable value on a given class is independent of the values of the other variables. This assumption is called as class conditional independence (Flach and Lachiche, 2004). Although in reality, the class conditional independence assumption is not always accurate, still surprisingly, naïve Bayes classifier gives good results. The true reason is that its classification accuracy does not depend on the dependencies that may exist among the attributes (Domingos and Pazzani, 1997). It ignores the interaction between variables within individuals of the same class. An important advantage of the naïve Bayes classifier is that it requires a small size of training data to estimate the parameters necessary for classification (Wikipedia, November 23, 2009). Here the training sample size is only 35 which is relatively small and hence naïve Bayes classifier is used for the purpose of predicting the appropriate class of an incumbent all-rounder. The naïve Bayesian classification works in the following way:

i) Each data sample is considered as an *n*-dimensional feature vector, $\boldsymbol{X} = (X_1, X_2, ..., X_n)$. This actually is the values of the *n* variables measured from the same sample member (subject) having *n* attributes A_1, A_2,..., A_n respectively.

ii) Suppose there are *m* classes, $C_1, C_2,..., C_m$ to which the subjects are to be classified. Given an unknown sample vector $\boldsymbol{X}$, the classifier will predict to which class the sample vector $\boldsymbol{X}$ is supposed to belong using the highest posterior probability, conditioned by $\boldsymbol{X}$. Thus, in this classifier, an unknown sample vector $\boldsymbol{X}$ is classified to the class C_i if and only if the probability of the sample vector $\boldsymbol{X}$ belonging to the i^{th} class C_i i.e. $P(C_i|\boldsymbol{X})$ is more than any class other than the i^{th} class. See equation (G-1) in Appendix G.

iii) The expression $P(C_i|\boldsymbol{X})$ is called as the posterior probability of the class C_i given the sample vector $\boldsymbol{X}$. Thus, one has to find C_i that maximizes the posterior probability $P(C_i|\boldsymbol{X})$. From Bayes theorem we can get an expression for $P(C_i|\boldsymbol{X})$. See equation (G-2) in Appendix G.

iv) As the denominator of the expression of posterior probability $P(C_i|\boldsymbol{X})$ is constant irrespective of the value of *i*, so maximizing of $P(C_i|\boldsymbol{X})$ is equivalent to maximization of the numerator only i.e. $P(\mathbf{X}|C_i)P(C_i)$. See equation (G-2). The term $P(C_i)$ is called as the class prior probability of the i^{th} class.

v) The attributes are assumed to be independent of each other. The independence of the attributes under consideration viz. strike rate in ODI, economy rate in ODI, strike rate in Twenty-20, economy rate in Twenty-20, and bowling type are tested. It is found

that the independence of the attributes between economy rate in ODI and economy rate in Twenty-20 is not factual (see Appendix H) while others are. However, the classification accuracy of naïve Bayesian classifier does not depend on the dependencies that exist among the attributes (Domingos and Pazzani, 1997). In case of independence of attributes, the expression $P(X|C_i)\ P(C_i)$ can be estimated from the training sample. Details are provided in Appendix G.

vi) Once the values of $P(X|C_i)\ P(C_i)$ be determined for each value of i, sample $\boldsymbol{X}$ is then assigned a class C_i if $P(X|C_i)\ P(C_i) > P(\mathbf{X}\ |C_j)P(C_j)\forall j \neq i$

The success of this classifier depends on the accuracy of the assumptions, viz., independence of the attributes and normality of the attributes that are continuous in nature.

Computation of Prior and Posterior Probabilities

The calculations are based on the values of the significant independent variables, viz., strike rate of ODI cricket (X_1), strike rate of Twenty-20 cricket (X_2), economy rate of ODI cricket (X_3), economy rate of Twenty-20 cricket (X_4), and bowling type (X_5), either spin or fast, obtained from Appendix A. The results of such probabilities are revealed in Appendix D and Appendix E.

However, before considering the continuous variables to be normally distributed, it is essential to check this normality assumption through appropriate normality test. Here the Kolmogorov-Smirnov test is used to check the normality assumption of the variables with parameters estimated from data (See Appendix C). The assumption of independence of the attributes is tested in Appendix H. It is seen that the independence of the attributes between economy rate in ODI and economy rate in Twenty-20 is not accurate while others are. In case the normality assumption is violated, then some algebraic, exponential, logarithmic transformation may be performed to attain normality for the corresponding variable. Based on the probabilities in Appendix D and Appendix E, one can determine the class (C_i) with maximum posterior probability for a given sample vector $\boldsymbol{X}$. The model would help to classify several such subjects into an appropriate class, viz., performer, batting all-rounder, bowling all-rounder and under-performer.

External Validation of the Model

To test the external validity of the model, a different sample of all-rounders is considered. The players considered for the validity exercise had not participated in the first three versions of IPL but had played in the fourth season only. To get sufficient information about their performance in IPL IV, the players who played at least 5 innings and bowled at least 10 overs are considered. Six all-rounders are found to satisfy the above criteria and they are – J Franklin, N McCullum, S Randiv, SA Hasan, T Southee, and W Parnell. Information about these players for the significant variables, viz., strike rate in ODI cricket, economy rate in ODI cricket, strike rate in Twenty-20 cricket, economy rate in Twenty-20 cricket, and bowling type are collected from *www.crickinfo.org* on April 5, 2011 (See Appendix F). For each all-rounder, the values of the independent variables are replaced in (G-4) (See Appendix-G) and using the appropriate distribution under each class, the posterior probability that the player would plummet in a given class is computed using (G-2) (See Appendix- G). This is repeated for each of the classes for the same player. The class with maximum probability would be the expected one for that player. The performances of the six all-rounders in IPL IV are recorded and their actual category is determined. The expected results from the classification model and the actual class based on performance from IPL IV are provided in Table 3 and it can be seen that the model can predict four of the six cases correctly and in one case (Randiv, S), the model missed the correct prediction narrowly.

Thus, depending on the team's requirement, a franchisee can decide which player to bid for. If the team requires a bowling all-rounder, then N McCullum, SA Hasan, and W Parnell are the options available. All the franchisees would definitely prefer not to bid for T Southee because he fell in the under-performer category as per the findings of the aforesaid model.

The fresh auction of IPL IV was held on January 8 and 9, 2011 with two new teams joining the league. IPL IV auctions saw an enormous rise in the players' price compared to the auctions of IPL I. The average bid price of the IPL IV auction is $495,605 while in the previous auction (before IPL I), it was $448,217. As evident from Table 3, all the incumbent all-rounders were sold at a price below the average. This practically does not reflect the importance of all-rounders in a cricket team especially for the above-mentioned players.

Table 3: Estimated Class of Six Incumbent All-rounders in Indian Premier League

No.	Player's Name	Class ofAll-rounders	Posterior Probability	Expected Class	Performance in IPL IV	Bid Amount
1	Franklin, J	Performer	0.06622	Bowling all-rounder	Batting all-rounder	$100,000
		Batting all-rounder	0.00232			
		Bowling all-rounder	**0.54450**			
		Under-performer	0.38694			
2	McCullum, N[3]	Performer	0.09938	Bowling all-rounder	Bowling all-rounder	$100,000
		Batting all-rounder	0.01705			
		Bowling all-rounder	**0.83634**			
		Under-performer	0.04721			
3	Randiv, S	Performer	**0.54721**	Performer	Bowling all-rounder	$80,000
		Batting all-rounder	0.00045			
		Bowling all-rounder	0.43736			
		Under-performer	0.01496			
4	Hasan, SA[3]	Performer	0.10416	Bowling all-rounder	Bowling all-rounder	$425,000
		Batting all-rounder	0.00077			
		Bowling all-rounder	**0.87107**			
		Under-performer	0.02398			
5	Southee, T[3]	Performer	0.05431	Under performer	Under- performer	$100,000
		Batting all-rounder	0.00786			
		Bowling all-rounder	0.19475			
		Under performer	**0.74307**			
6	Parnell, W[3]	Performer	0.10284	Bowling all-rounder	Bowling all-rounder	$160,000
		Batting all-rounder	0.00001			
		Bowling all-rounder	**0.51083**			
		Under-performer	0.38630			

CONCLUSION

IPL is a young professional league; yet it has planned its player distribution procedure in the same lines as of the other professional sports leagues around the world. The salaries of players that are decided through auction are a way of quantifying players' performance in monetary terms. Thus, it is a matter of decision making on the part of the franchisee to determine which player to bid for and up to what cost. This model can help a franchisee to take such decisions. The salaries offered to the players made prior to the first season of IPL were valid for three years. At the beginning of the fourth season of IPL, new agreements are made through fresh bidding. If the league continues, fresh bidding shall take place after every three years; and in that context, this study may be used to predict the probable category of all-rounders. This prediction can help the franchisee to decide which all-rounders should be targeted for their team and who should not be considered.

The model when applied for external validity was found to be 66.7 per cent accurate. This could have been the result of the fact that the training sample used to develop the model was loaded with only 35 all-rounders. As the age of the league increases, more number of all-rounders can be considered in the training sample and much better predictions could be expected from the model. However, the form of the player at the time of the actual tournament is a deterministic factor for the actual classification. But the model once rich in data, is supposed to work well provided the performance of the player is not much different from his performance in the previous matches. The IPL is likely to be around for many years, providing several opportunities for further research (Parker, Burns and Natarajan, 2008).

[3] The predicted and actual performance/categories of these players are similar by applying naïve Bayesian model.

Appendix A: Brief Description of Different Independent Variables

Players' Name	Age	L/R bat (1-R, 0-L)	Bowl Type (1-fast, 2-spin)	No. of T20 Matches Played	No. of ODI Matches Played	SR in ODI Cricket	SR in T20 Cricket	Econ. Rate in ODI	Econ. Rate in T20	Years of Inter-national Cricket	Country	Avg. Batting Position in IPL	L/R Bowl (1-R, 0-L)	Bidding Amount in IPL1
Morkel, JA	29	0	1	31	51	100.1	142.9	5.46	8.03	7	SA	5.13	1	675,000
Oram, JDP	32	0	1	27	151	85.6	138.46	4.35	8.74	10	NZ	7.47	1	675,000
Shahid Afridi	31	1	2	43	325	113.92	144.09	4.6	6.21	14	PAK	4.22	1	675,000
Raina, SK	24	0	2	19	115	89.83	149.99	5.35	9.75	6	IND	1.83	1	650,000
Chawla, PP	22	1	2	3	25	65.51	112.92	5.1	6.27	5	IND	7.25	1	400,000
Sharma, RG	24	1	2	20	61	75.72	128.9	4.83	7.4	4	IND	3.94	1	750,000
Styris, SB	35	1	1	31	188	79.41	119.66	4.74	6.77	9	NZ	4.85	1	175,000
Symonds, A	35	1	2	14	128	92.44	169.34	5	8.98	13	AUS	3.8	1	13500,000
Dilshan, TM	34	1	2	32	203	87.54	120.7	4.73	8.94	11	SL	3.08	1	250,000
Maharoof, MF	26	1	1	7	94	85.19	85.18	4.78	7.2	10	SL	8.08	1	225,000
Sehwag, V	32	1	2	14	236	104.1	153.43	5.3	20	12	IND	1.52	1	833,750
Agarkar, AB	33	1	1	4	191	80.62	136.36	5.1	8.09	9	IND	6.33	1	350,000
Ganguly, SC	36	0	1	58*	311	73.7	109.74	5.1	7.73	16	IND	3.12	1	1092,500
Gayle, CH	31	0	2	20	228	83.95	144.49	4.73	7.29	12	WI	1.76	1	800,000
Hodge, BJ	33	1	2	8	25	87.51	122.07	4.63	10	4	AUS	3.22	1	30,000
Hussey, DJ	33	1	2	28	34	89.94	127.19	5.3	6.36	4	AUS	4.97	1	625,000
Shukla, LR	29	1	1	42*	3	94.73	128.41	4.94	7.68	11	IND	5.79	1	30,000
Hopes, JR	32	1	1	12	84	93.71	107.14	4.53	7.64	6	AUS	3.38	1	30,000
Pathan, IK	26	0	1	16	107	77.68	125.47	5.25	7.74	8	IND	5.5	0	925,000
Yuvraj Singh	29	0	2	23	274	87.58	151.6	5.04	8.08	11	IND	4.08	0	1063,750
Bravo, DJ	27	1	1	22	113	81.93	123.29	5.24	8.71	7	WI	5.38	1	1000,000
Duminy, JP	27	0	2	30	78	85.07	126.31	5.02	8.66	7	SA	5.11	0	950,000
Jayasuriya, ST	40	0	2	30	444	91.22	129.64	4.78	7.44	21	SL	1.63	0	975,000
Pollock, SM	36	1	1	12	303	86.69	122.85	3.67	7.62	13	SA	5.87	1	550,000
Kallis, JH	35	1	1	16	314	72.77	119.9	4.81	7.38	16	SA	2.05	1	900,000
Kohli, V	22	1	1	3	54	82.13	135	6.17	8.48	3	IND	3.88	1	30,000
Kumar, P	24	1	1	4	48	85.22	60	5.07	5.44	4	IND	8.14	1	300,000
Jadeja, RA	22	0	2	9	35	76.97	86.66	4.86	7.48	2	IND	5.68	0	30,000
Pathan, YK	28	1	2	19	51	115.14	147.55	5.46	8.7	4	IND	5	1	475,000
Warne, SK	38	1	2	53*	194	72.04	95.85	4.25	7.23	15	AUS	9.2	1	450,000
Watson, SR	29	1	1	22	133	89.59	148.69	4.85	7.65	9	AUS	2.32	1	125,000
Ryder, JD	26	0	1	17	35	90.99	122.22	6.25	6.8	3	NZ	1	1	160,000
Vettori, DL	32	0	2	28	272	81.93	109.35	4.12	5.36	13	NZ	6.47	0	225,000
Pollard, KA	24	1	1	20	39	103.12	124.18	5.28	8.37	4	WI	5.11	1	750,000
Mathews, AD	23	1	1	21	43	85.13	126.29	4.65	7.33	2	SL	5.14	1	950,000

* Did not play International Twenty-20 matches till April 2011. The data from other Twenty-20 matches are used.

Source: www.cricinfo.org

Appendix B: Performance of the All-rounders up to IPL3

No.	Players' Name	IPL Career			Country	IPL Team
		Matches	Strike Rate	Economy Rate		
1	Morkel, JA	25	145.06	8.26	SA	CSK
2	Oram, JDP	15	100.95	8.95	NZ	CSK
3	Shahid Afridi	10	176.09	7.5	PAK	DC
4	Raina, SK	30	141.56	6.26	IND	CSK
5	Chawla, PP	29	127.37	7.6	IND	KXIP
6	Sharma, RG	29	130.27	6.89	IND	DC
7	Styris, SB	10	96.18	7.51	NZ	DC
8	Symonds, A	12	150.18	8.46	AUS	DC
9	Dilshan, TM	27	127.16	8.93	SL	DC
10	Maharoof, MF	21	124.29	5.94	SL	DD
11	Sehwag, V	13	146	6.83	IND	DD
12	Agarkar, AB	25	168.72	13.17	IND	DD
13	Ganguly, SC	20	116.5	8.61	IND	KKR
14	Gayle, CH	26	104.67	7.03	WI	KKR
15	Hodge, BJ	7	119.58	7.65	AUS	KKR
16	Hussey, DJ	15	115.29	8.05	AUS	KKR
17	Shukla, LR	17	131.13	9.65	IND	KKR
18	Hopes, JR	22	118.84	8.62	AUS	KKR
19	Pathan, IK	11	149.32	9.86	IND	KXIP
20	Yuvraj Singh	28	119.34	7.15	IND	KXIP
21	Bravo, DJ	29	133.4	7.76	WI	KXIP
22	Duminy, JP	20	117.51	7.98	SA	MI
23	Jayasuriya, ST	13	114.46	5.31	SL	MI
24	Pollock, SM	26	146.71	8.05	SA	MI
25	Kallis, JH	27	137.59	8	SA	MI
26	Kohli, V	13	132.43	6.54	IND	MI
27	Kumar, P	26	108.74	8.27	IND	RCB
28	Jadeja, RA	29	109.31	7.83	IND	RCB
29	Pathan, YK	26	122.92	8.06	IND	RCB
30	Warne, SK	27	116.53	6.75	AUS	RR
31	Watson, SR	29	159.15	7.49	AUS	RR
32	Ryder, JD	28	101.71	7.54	NZ	RR
33	Vettori, DL	15	151.77	7.07	NZ	RR
34	Pollard, KA	5	114.29	6.76	WI	RCB
35	Mathews, AD	9	106.25	7.54	SL	DD

Source: www.cricinfo.org

Appendix C: Normality Check for Four Continuous Variables: One-Sample Kolmogorov-Smirnov Test

		Strike Rate of ODI	Strike Rate of Twenty-20	Economy Rate of ODI	Economy Rate of Twenty-20
Most Extreme Differences	Absolute	0.124	0.163	0.099	0.258
	Positive	0.124	0.083	0.099	0.258
	Negative	0-.060	-0.163	-0.099	-0.159
Kolmogorov-Smirnov Z		0.733	0.965	0.588	1.528
Asymp. Sig. (2-tailed)		0.656	0.309	0.879	0.119
The test distribution is		Normal	Normal	Normal	Normal

Appendix D: Prior Class Probabilities Derived from the Data in Appendix A

Class	Notation of Class Probability	Class Probabilities
Performer	$P(C_1)$	0.257
Batting All-rounder	$P(C_2)$	0.229
Bowling All-rounder	$P(C_3)$	0.286
Under-performer	$P(C_4)$	0.229

Appendix E: Distribution of the Significant Variables under Different Classes

Variable Class	SR of ODI$N(\mu,\sigma^2)$	SR of Twenty-20$N(\mu,\sigma^2)$	ER of ODI$N(\mu,\sigma^2)$	ER of Twenty-20$N(\mu,\sigma^2)$	Bowling Type	
					Spin	Fast
Performer	μ = 91.634 σ = 16.546	μ = 129.372 σ = 21.327	μ = 4.88 σ = 0.539	μ = 7.686 σ = 1.133	0.556	0.444
Batting All-rounder	μ = 93.028 σ = 6.305	μ = 138.441 σ = 19.654	μ = 5.008 σ = 0.335	μ = 9.233 σ = 4.414	0.625	0.375
Bowling All-rounder	μ = 80.928 σ = 6.074	μ = 116.045 σ = 16.439	μ = 4.905 σ = 0.588	μ = 7.4 σ = 1.008	0.6	0.4
Under-performer	μ = 83.814 σ = 6.285	μ = 120.436 σ = 25.405	μ = 5.039 σ = 0.538	μ = 8.065 σ = 1.325	0.125	0.875

Appendix-F: Data Related to Incumbent All-rounders

No	Player's Name	SR in ODI	SR in Twenty-20	ER in ODI	ER in Twenty-20	Bowling Type
1	Franklin	78.64	100.52	5.15	7.23	Fast
2	McCullum, N	82.68	106.96	4.64	6.32	Spin
3	Randiv, S	67.82	133.33	4.66	6.76	Spin
4	Hasan, SA	76.52	111.89	4.29	6.69	Spin
5	Southee, T	85.26	89.28	5.23	8.5	Fast
6	Parnell, W	70.73	87.5	5.96	7.02	Fast

Source: www.cricinfo.org

Appendix G: Equations Related to Naïve Bayesian Classification

A sample vector $\boldsymbol{X}$ is classified to a class C_i if and only if

$$P(C_i|\boldsymbol{X}) > P(C_j|\boldsymbol{X}) \text{ for } 1 \le j \le m \text{ and } j \ne i \quad \text{(G-1)}$$

From Bayes theorem we can get an expression for $P(C_i|\boldsymbol{X})$

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{\sum P(\mathbf{X}|C_i)P(C_i)} \quad \text{(G-2)}$$

The class prior probability would be estimated by (G-3), i.e.,

$$P(C_i) = \frac{s_i}{s} \quad \text{(G-3)}$$

where, s_i is the number of training samples of class C_i, and s is the total number of training samples.

The attributes are assumed to be independent of each other. Then one can write

$$P(\mathbf{X}|C_i)P(C_i) = \prod_{k=1}^{n} P(X_k|C_i)P(C_i) \quad \text{(G-4)}$$

Now, the probabilities $P(X_k|C_i)$ $(k = 1, 2,\ldots,n)$ can be estimated from the training sample in the following manner:

a) If the k^{th} attribute A_k, is categorical, then

$$P(X_k | C_i) = \frac{s_{ik}}{s_i} \quad \text{(G-5)}$$

where, s_{ik} in the number of training sample points of class C_i for which the k^{th} attribute attains the value X_k.

b) However, if the k^{th} attribute A_k, is continuous, then $P(X_k | C_i) \sim N(\mu_{c_i}, \sigma_{c_i})$, so that

$$p(x_k \mid c_i) = \frac{1}{\sqrt{2\pi}\,\sigma_{c_i}} e^{-\frac{(x_k - \mu_{c_i})^2}{2\sigma_{c_i}^2}} \quad \text{(G-6)}$$

where, are the mean and standard deviation of the values of the attribute A_k belonging to the i^{th} class. However, before setting up the distribution it is obligatory to test such a claim. In case of non-normality the appropriate distribution be identified and its parameters be estimated from the data provided by the training sample and then to proceed in the same manner as in this case.

Appendix H: Test of Independence of the Attributes under Consideration

The median test can be used to check the independence between two samples when both are continuous in nature. At the very outset each of the data sets, viz. economy rate in ODI (ER ODI), economy rate of Twenty-20 (ER T20), strike rate in ODI (SR ODI) and strike rate in Twenty-20 (SR T20) are standardized in order to make them comparable. To test the independence between two samples of sizes n_1 and n_2 (say), the samples are pooled and the median of the combined sample is determined. In this case $n_1 = n_2$. The number of observations for each sample that is above and below the combined median is counted. Let the numbers be m_1 and m_2. The null hypothesis is that the samples are independent of each other. The joint distribution of m_1 and m_2 under the null hypothesis follows hyper geometric distribution, i.e.,

$$P(m_1, m_2) = \frac{{}^{n_1}C_{m_1} \times {}^{n_2}C_{m_2}}{{}^{n_1+n_2}C_{m_1+m_2}} \quad \text{(I-1)}$$

Now, using the above distribution of m_1 and m_2, we can perform χ^2 test with 1 d.f. to test the null hypothesis of independence provided the values of n_1 and n_2 are sufficiently large. The test statistic of which is given by,

$$\chi^2 = \frac{[m_1(n_2 - m_2) - m_2(n_1 - m_1)]^2 N}{m_1(n_2 - m_2)m_2(n_1 - m_1)} \sim \chi^2 \text{ with 1 d.f.} \quad \text{(I-2)}$$

where, N is the total frequency.

Median Test

SR ODI	SR T20	ER ODI	ER T20
SR ODI	3.44 (*0.063*)	0.252 (*0.615*)	3.44 (*0.063*)
SR T20		1.39 (*0.238*)	3.44 (*0.063*)
ER ODI			10.3 (*0.001*)

Note: The values in parenthesis are the corresponding *p*-values of the chi-square test.

From the above table it is seen that except economy rate in ODI and economy rate in Twenty-20 all other attributes are independent of each other.

Test of Independence of Bowling Type with Other Attributes

The Wald Wolfowitz Run test, also called as two sample run test, is used to check whether the two groups have similarity in the attribute under consideration. The test can be used to check if a cricketing attribute say economy rate in ODI is independent of the type of bowling viz. spin or fast. To perform this test economy rate of fast bowlers and spinners are arranged in ascending order of magnitude and then the ordered sample is rewritten as a sequence of 1 and 2 such that, 1 represents a fast bowler and 2 represents spin bowler. Let K be the total number of runs in the sequence of 1 and 2. So,

if both the samples are independent of each other then there will be a thorough mixing of 1's and 2's and accordingly the number of runs will be more. However, if the significance value of corresponding variable is less than 0.05 then two samples are considered to be independent.

Wald Wolfowitz Run Test

	Bowling Type (fast=1, spin=2)			
	No. of Runs	***Z***	**Asymp. Sig. (1-tailed)**	**Comments**
SR ODI	15	-1.025	0.013	Bowling type is independent of all other attributes
SR T20	16	-0.682	0.048	
ER ODI	21	1.035	0.036	
ER T20	13	-1.712	0.043	

Since the *p*-value corresponding to all the variables are less than 0.05, therefore the independence of the variates with the attribute, viz., type of bowling holds.

REFERENCES

Agresti, A and Finley, B (1986). *Statistical Methods for the Social Sciences*, Second Edition, San Francisco: Dellen.

Bailey, T (1989). *The Greatest Since My Time*, London: Hodder and Stoughton.

Bloomfield, J; Polman, R and O'Donoghue, P (2007). "Physical Demands of Different Positions in FA Premier League Soccer," *Journal of Sports Science and Medicine*, 6, 63-70.

Bracewell, P J and Ruggiero, K (2009). "A Parametric Control Chart for Monitoring Individual Batting Performances in Cricket," *Journal of Quantitative Analysis in Sports*, 5(3), 1-19.

Brettenny, W (2010). *Integer Optimization for the Selection of a Fantasy League Cricket Team,* Unpublished M.Sc Dissertation, Faculty of Science, Nelson Mandela Metropolitan University, South Africa.

Cardus, N (1978). *Cardus in the Covers*, United Kingdom: MacDonald and Company Limited.

Carvajal, W; Ríos, A; Echevarría, I; Martínez, M; Miñoso, J and Rodríguez, D (2009). "Body Type and Performance of Elite Cuban Baseball Players," *MEDICC Review*, Spring, 11(2), 15-20.

Clerke, S R (1997). *Performance Modeling in Sports*, Unpublished Ph.D dissertation, Submitted to the School of Mathematical Sciences, Swinburne University of Technology.

Cox, D R and Snell, D J (1989). *The Analysis of Binary Data*, Second Edition, London: Chapman and Hall.

Das, S P (2008). "Game of Organizing International Cricket: Co-existence of Country-line and Club-line Games," *Economics: The Open-Access, Open-Assessment E-Journal*, 2(32), 1-29.

Dwivedi, L K; Ram, F and Reshmi, R S (2007). "An Approach to Understand Change in Contraceptive Behavior in India," *Genus,* LXIII(3 & 4), 19-54.

Domingos, P and Pazzani, M (1997). "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss Function," *Machine Learning*, 29(2-3), 103-130.

Field, A (2009). *Discovering Statistics using SPSS*, London: Sage Publications Ltd., 269.

Flach, P A and Lachiche, N (2004). "Naïve Bayesian Classification of Structured Data," *Machine Learning*, Boson: *Kluwer Academic Publishers*, 1-37.

Fraboni, R and Rosina, A (2006). "Age at First Union and Fatherhood in a Very Low Fertility Context," *Genus,* LXII(3 & 4), 87-109.

Gabbett, T J (2002). "Physiological Characteristics of Junior and Senior Rugby League Players," *British Journal of Sports Medicine*, 36, 334-339.

Hosmer, D W and Lemeshow, S (1989). *Applied Logistic Regression*, New York: Wiley.

Indongo, N and Naidoo, K (2009). "Family Planning Dialogue: Identifying the Key Determinants of Young Women's Use and Selection of Contraception in Namibia," *Demography India,* 38(1), 17-34.

Knight, J (2006). *Cricket for Dummies,* John Wiley & Sons Ltd., Chichester, West Sussex, England, 156.

Lemmer, H (2004). "A Measure for the Batting Performance of Cricket Players," *South African Journal for Research in Sport, Physical Education and Recreation*, 26(1), 55-64.

Lemmer, H (2006). "A Measure of the Current Bowling Performance in Cricket," *South African Journal for Research in Sport, Physical Education and Recreation*, 28(2), 91-103.

Lemmer, 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.

Lenten L J A; Geerling, W and Kónya, L (2009). "A Hedonic Model of Player Wage Determination from the Indian Premier League Auction" Seminar Paper, Department of Economics and Finance, La Trobe University, Australia. Available at *www.econ.canterbury.ac.nz/research/pdf/Paper_Lenten.pdf, Accessed on October 15, 2010.*

Lewis, A J (2005). "Towards Fairer Measures of Player Performance in One-Day Cricket," *The Journal of the Operational Research Society*, 56(7), 804- 815.

Matsumoto, Y and Thawonmas, R (2004). "MMOG Player Classification Using Hidden Markov Models," *Lecture Notes in Computer Science*, 3166, 15-20.

McGee, K J and Burkett, L N (2003) "The National Football League Combine: A Reliable Predictor of Draft Status?" *Journal of Strength and Conditioning Research*, 17(1), 6-11.

Menard, S (1995). *Applied Logistic Regression Analysis*, Sage University Paper Series on Quantitative Application on Social Sciences, Thousand Oaks, CA: Sage, 07-106.

Nagelkerke, N J D (1991). "A Note on a General Definition on the Coefficient of Determination," *Biometrika*, 78(3), 691-692.

Njogu, W and Martin, T C (2006). "The Persisting Gap Between HIV/AIDS Knowledge and Risk Prevention Among Kenyan Youth," *Genus*, LXII(2), 135-168.

Parker, D; Burns, P and Natarajan, H (2008) "Player Valuations in the Indian Premier League," *Frontier Economics*, October, 1-17.

Ramani, S (2008) "Cricket, Excesses and Market Mania," *Economic and Political Weekly*, March 8, 13-15.

Rastogi, S K and Deodhar, S Y (2009). "Player Pricing and Valuation of Cricketing Attributes: Exploring the IPL Twenty-20 Vision," *Vikalpa*, 34(2), 15-23.

Rundell, M (2006). *The Wisden Dictionary of Cricket*, Third Edition, London: A and C Black Publishers Ltd., 1-2.

Schwandt, M J; Glenn, J R and Triantis, K (2007). Intercollegiate Athletic Department's Performance Assessment, *Conference Proceedings the 2007 International Conference of the System Dynamics Society*, Boston, Massachusetts, USA.

Singh, S; Gupta, S and Gupta, V (2011). "Dynamic Bidding Strategy for Players in IPL," *International Journal of Sports Science and Engineering*, 5(1), 3-16.

Tan, A and Ramachandran, R (2010). "Ranking the Greatest All-rounders in Test Cricket," Available at *www.cricket society.com/ranking_the_greatest_all-ro.pdf*, Accessed on October 15, 2010.

Tabachnick, B G and Fidell, L S (2001). *Using Multivariate Statistics*, 4th edition, Boston: Allyn & Bacon.

Vig, A (2008). "Efficiency of Sports League – The Economic Implications of Having Two in the Indian Cricket Market," Masters Project: The University of Nottingham.

Van Staden, P J (2008). "Comparison of Bowlers, Batsmen and All-rounders in Cricket Using Graphical Display," Technical Report 08/01, Department of Statistics, University of Pretoria, South Africa.

Van Staden, P J (2009). "Comparison of Cricketers' Bowling and Batting Performances Using Graphical Displays," *Current Science*, 96(6), 764-766.

Wikipedia (2009). "Naïve Bayes Classifier," Available at URL *http://en.wikipedia.org/wiki/Naive_Bayes_classifier*, Accessed on November 23, 2009.

Zhang, H (2004). "The Optimality of Naïve Bayes," *American Association for Artificial Intelligence*, Available at *http://* www.aaai.org, Accessed on 19th November, 2010.

Hemanta Saikia is a Research Scholar in the Department of Business Administration, Assam University, Silchar, Assam. He has completed his Masters in Statistics from Gauhati University in 2007 and is currently perusing his Ph.D in Data Mining in Sports.

e-mail: h.saikia456@gmail.com

Dibyojyoti Bhattacharjee is a Reader in the Department of Business Administration, Assam University, Silchar. He has an M.Sc and an M. Phil. in Statistics from University of Delhi. His Ph.D is in Statistical Graphics, completed from the Department of Statistics, Gauhati University. He has to his credit 30 research publications in different national and international journals and thirteen books in different sub-fields of Statistics. He has served various other institutes like G C College, Ramlal Anand College, Central Statistical Organization, Gauhati University, etc. He was awarded with fellowships from ICSSR and UGC. He is a member on the Editorial Board of *Interstat* and *ISST Journal of Applied Mathematics* and the Managing Editor of the *Journal of Empirical Research* in *Social Science*.

e-mail: djb.stat@gmail.com