Paper 019-2013

Predictive Modeling in Sports Leagues: An Application in Indian Premier League

Pankush Kalgotra, Ramesh Sharda, Goutam Chakraborty, Oklahoma State University, OK, US

ABSTRACT

The purpose of this article is to develop models that can help team selectors build talented teams with minimum possible spending. In this study, we build several predictive models for predicting the selection of a player in the Indian Premier League, a cricket league, based on each player's past performance. The models are developed using SAS® Enterprise Miner™ 7.1. The best-performing model in the study is selected based on the validation data misclassification rate. The selected model provides us with the probability measure of the selection of each player, which can be used as a valuation factor in the bidding equation. The models that are developed can help decision-makers during auction set salaries for the players.

INTRODUCTION

Team selection is a highly critical process in every sport as players are selected based on their past performance. Forecasting future from the past is highly subjective and thus requires extraordinarily expert decision-making. It becomes more prominent when a huge amount of money is involved. In this paper, we build several predictive models for the selection of players based on their past performances using data mining techniques and tools. We use data from one of the most famous professional sports leagues in the world known as the Indian Premier League, a cricket tournament. The models provide the probability of selection of each player in the team, which can be used as a valuation factor in linear cost function for estimating the players' bid amount.

Cricket, the second most popular sport in the world after soccer, is a bat-and-ball game played between two teams of 11 players on a field, at the centre of a 22-yard long rectangular pitch. There are three different forms of cricket - Test, One-day and Twenty20. Test cricket is the traditional and longest form of cricket in which a match can continue up to maximum five days or maximum 90 overs (six bowls are bowled in an over) a day. Four innings are played in a match with each team batting twice. The second form is the One-day cricket, which is played over one day; each team bats maximum 50 overs. The latest form of cricket is the Twenty20 or T20 cricket. It involves two teams with each team batting maximum of 120 balls i.e. twenty overs and is completed in an about two and half hours, a much shorter duration. The England and Wales Cricket Board (ECB) in England originally introduced Twenty20 for professional inter-county competition in 2003 but it is now famous all around the world. This form of cricket has attracted large crowds to the stadiums and viewers on televisions because of its fast-paced nature and intense competition. Unexpected results of the games make it even more attractive (Beaudoin, 2003). The International Cricket Council organizes a world cup tournament every four years for One-day cricket; it also organizes a world cup tournament for Twenty20 cricket every two years.

Indian Premier League (IPL) is one of the famous Twenty20 cricketing events and takes place every year in India (although second season in 2009 was hosted by South Africa because of security issues in India). The Board of Control for Cricket in India (BCCI) initiated IPL in 2008. It is the first sporting event ever to be broadcast live on YouTube (in association with Indiatimes in 2010). According to the inaugural annual review of Global Sports Salaries published by sportingintelligence.com in April 2010, the IPL is the second highest-paid league in the world after the National Basketball Association (NBA), based on first-team salaries on a pro-rata basis. It has become second highest paid sports league in the world within three years of its inauguration. Its brand value was estimated to be around $2.99 billion in the fifth season (2012) reported by Brand Finance (a brand valuation consultancy). In the first season of IPL, eight teams played amongst each other and later in 2011, two more teams were introduced in the contest. In 2012 season, one of the two new teams violated the terms and agreements with the BCCI and therefore, was not allowed to play. See Table 1 for the team names and their owners.

In 2008, the IPL became the first cricketing event in which players were bought through auctions. Cricketers around the world register themselves for the auction pool with or without their base price set by BCCI (Players were allowed to set their base price between $200,000 and $400,000 in 2011 and 2012 seasons). Franchise or the team owners cannot bid for a player below his base price. A common limit is set for every franchise to spend at the IPL auction so that rich owners could not buy the entire list of best players and destroy the semblance of the contest (Zimbalist, 2002). Franchise bid on the players from the pool consist of batsmen, bowlers and all-rounders. There are five different ways that a franchise can acquire a player following team composition rules (See appendix 1 for team composition rules). In the annual auction, buying domestic players, signing uncapped players, through trading, and by buying replacements. In the annual auction, highest bidder on a player signs a fixed three-year contract with him. It is an extremely difficult decision for a franchise owner to buy a talented team with a limited amount of money.

Past studies have shown that pay cannot be adequately explained by past performance alone, nor are pay levels justified by future performance in IPL (Dalmia, 2010). Further evidence was found by Karnik (2009) that more expensive players often provide a lower rate of return to the owners, which brings into question the validity of bidding up a player at auction. This makes very hard for the franchise to spend money on the players efficiently. So, the franchisee owners need a robust model that can be used to estimate the player's bid amount.

In this paper, we constructed several predictive models to get the probability of selection of each player that may be used as a valuation factor in the cost function (cost function can vary from one franchise to other) for estimating his bid amount. Data from the Twenty20 cricket including first four seasons (2008-2011) of the IPL was used as the input to the model to predict the selection of each player in the fifth season (2012). The rest of this paper is organized as follows. Section 2 provides a brief literature review. Section 3 describes the methodology. Section 4 presents results and section 5 concludes the paper by presenting the future work.

Team Name	Owner(s)
Mumbai Indians	Mukesh Ambani (Owner of Reliance Industries)
RoyalChallengers Bangalore	Vijay Malya (UB Group)
Hyderabad Deccan Chargers	T.Venkatram Reddy (Deccan Chronicle group)
Chennai Super Kings	India Cements, Gurunath Meiyappan (public face)
Delhi Daredevils	GMR Group
Kings XI Punjab	Ness Wadia, Preity Zinta, Dabur, Apeejay Surendera Group
Kolkata Knight Riders	Shahrukh Khan (Red Chillies Entertainment), Juhi Chawla, Jay Mehta
Rajasthan Royals	Emerging Media (Lachlan Murdoch), Shilpa Shetty, Raj Kundra
Pune Warriors (introduced in 2011)	Subrato Roy Sahara
Kochi TuskersKerela (introduced in 2011 and defunct in 2012)	Kochi Cricket Private Ltd

Table 1. IPL Teams and Owners (2008-Current)

LITERATURE REVIEW

A variety of studies have looked at performance and franchise bidding in the Indian Premier League. This paper is first ever study on the selection of players in the IPL teams and estimation of their bid amounts in the Twenty20 form of cricket. Following is a brief review of the work done on the IPL and related studies.

Parker, Burns and Natarajan (2008) explored the determinants of valuations and investigated a number of hypotheses related to the design of the auction in IPL using information of the previous performance, experience, and other characteristics of individual players.

Iyer and Sharda (2009) used neural networks in forecasting the selection of athletes in the cricket teams by predicting their future performance based on past performance. A prediction for the selection of a cricketer in the one-day international world cup 2007 was made. To predict the selection, players were categorized into a performer, a moderate or a failure.

Karnik (2009) followed a very simple approach to derive the hedonic price equations for estimating a bid amount for each cricketer in the Indian Premier League (IPL) auction. He developed price models using the data from the 2008 season and successfully tested against the data from the 2009 season. The variables used in the equations were the common playing factors such as runs scored, wickets taken and age. He observed a lower rate of return from the expensive players to the owners of the teams that showed the inefficiency in judging the pay levels of the players by the bidders.

Singh, Gupta and V. Gupta (2011) formulated an integer-programming model for the efficient bidding strategy for the franchises. The model was implemented in a spreadsheet that helped in taking bidding decisions in real time and overcome winner's curse, which is typically associated with normal bidding processes.

Singh (2011) made an effort to measure the performance of teams in the IPL using the non-parametric mathematical approach called Data Envelopment Analysis (DEA). He used both playing and non-playing factors for analyzing the efficiencies of the teams in 2009 season.

Lenten, Geerling, and Kónya (2012) analyzed various playing and non-playing factors of the athletes of cricket sport that determine their bid value in the auction of Indian Premier League. They considered every form of cricket to find the relationship between those factors and the wages of players. Several cross sectional models were estimated to find the combinational effects of the variables on the salary of the players. Most of the studies have focused on making the bidding decisions in the IPL but this paper is the one of the first studies focused on prediction of the selection of players in the IPL teams and estimation of their bid amounts. Due to the unavailability of sufficient data (league started in 2008), other scholars have used the data from other forms of cricket instead of Twenty20 but our models were built using data only from Twenty20 type of cricket, which ensured the relevancy of our models for this exciting and new form of cricket.

Variable	Level	Variable Description
All Rounder	Binary	Indicates whether is a player is an all-rounder or not. Value is 1 if batsman otherwise 0.
Innings	Interval	Number of innings played by the player instead of matches
Not out	Interval	Number of times a player has been not out in his career
Runs	Interval	Total number of runs in the Twenty20 career
HS	Interval	Highest runs in an innings by the player
Average	Interval	Total number of runs a player has scored divided by the number of times he is out
Country	Nominal	Players' country (12 dummy variables were created for the 13 countries)
Strike Rate	Interval	Average number of runs scored per 100 balls faced
Century	Interval	Number of hundreds in Twenty20 career
Half Century	Interval	Number of fifties in his Twenty20 career
4s	Interval	Number of fours in all the innings he has played in twenty20 so far.
6s	Interval	Number of sixes in all the innings
Catch	Interval	Number of catches taken by a player during fielding
Result (Target)	Binary	Selected in this IPL or not (1- selected and 0 - not selected)

Table 2. Batting Dataset

Variable	Level	Variable Description
Innings	Interval	Number of innings played by the player instead of matches
Balls	Interval	Number of balls bowled
Runs	Interval	Total number of runs conceded
Wkts	Interval	Number of wickets taken
Average	Interval	Average number of runs conceded per wicket
Country	Nominal	Players' country (12 dummy variables were created for the 13 countries)
Economy	Interval	Average number of runs conceded per over
SR	Interval	Average number of balls bowled per wicket taken
Best	Interval	Best innings bowling
4w	Interval	Number of innings in which the bowler took at least four wickets
5w	Interval	Number of innings in which the bowler took at least five wickets
Result (Target)	Binary	Selected in this IPL or not (1- selected and 0 - not selected)

Table 3. Bowling dataset

METHODOLOGY

We culled data from cricketarchive.com, www.iplt20.com, sports.ndtv.com and www.espncricinfo.com. Data consisted of all playing factors excluding fielding and wicket-keeping. These two abilities also contribute towards the performance of the players but the scope of this article is to predict the selection of batsmen and bowlers in the teams. We created two datasets; one for bowlers and another for batsmen. The players who can play a role of batsman and bowler both known as all-rounders and have a different role in the team. It would have been better if a different dataset for all-rounders is created but due to the small sample size, we included them in batsmen dataset with a variable named All-rounder that differentiated them from batsmen. The batsmen dataset contained a total of 13 independent variables and one dependent binary variable that shows whether a player is selected in the team or not. There were 11 input variables and one target variable in the bowling dataset. Twelve dummy variables were created for thirteen counties in both datasets to convert qualitative facts into numeric values. For complete information about all variables, see table 2 and table 3. Predictive modeling had been done in a sequence of steps known as SEMMA (Sample, Explore, Modify, Model and Assess) developed by SAS Institute Inc. as shown in the figure 1.

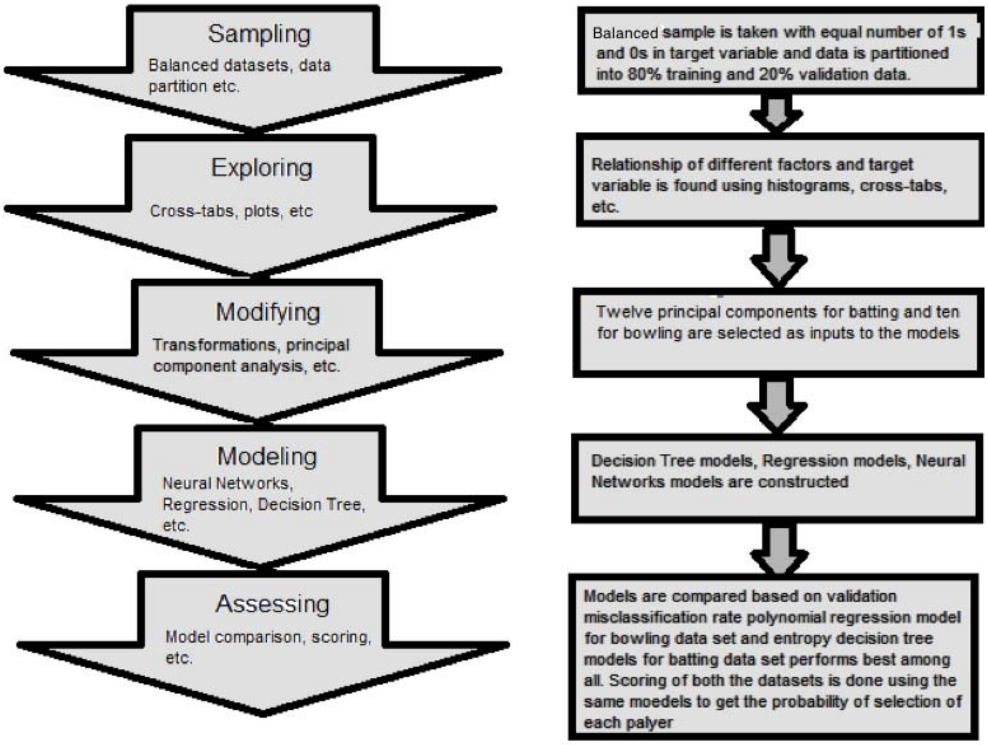
Sampling
Balanced datasets, data partition etc.
Exploring
Cross-tabs, plots, etc
Modifying
Transformations, principal component analysis, etc.
Modeling
Neural Networks, Regression, Decision Tree, etc.
Assessing
Model comparison, scoring, etc.
Balanced sample is taken with equal number of 1s and 0s in target variable and data is partitioned into 80% training and 20% validation data.
Relationship of different factors and target variable is found using histograms, cross-tabs, etc.
Twelve principal components for batting and ten for bowling are selected as inputs to the models
Decision Tree models, Regression models, Neural Networks models are constructed
Models are compared based on validation misclassification rate polynomial regression model for bowling data set and entropy decision tree models for batting data set performs best among all. Scoring of both the datasets is done using the same moedels to get the probability of selection of each palyer

Figure 1. SEMMA

We collected the data containing the information of 313 batsmen and 277 bowlers of different countries who have either played in any season of IPL or have ever registered themselves for the auction. These datasets contained the information of the players from their Twenty20 career including the four seasons of the IPL from 2008 to 2011. Using the data, prediction of the selection of players in the 2012 season was made. Following the SEMMA data mining approach, a sequence of steps was performed on both bowling and batting datasets for predictive modeling as shown in figure 1. After collecting the data, the next step was to create a balanced sample having equal number of 1s and 0s in the target variable. For adjusting the oversampling, prior probabilities had been set equal to the percentage of positive result in the datasets. Data was partitioned into 80% training data for the modeling purpose and 20% validation data for the validation purpose using stratified partitioning method.

Several exploratory tools such as crosstabs, histograms, pie charts, correlations, etc. were used to understand the relationship between target variable and other input variables. We found that the variable Country (Players' country) was strongly associated with the target variable and many input variables in both datasets. But, this is a structural relationship that happened to exist in the data due to nature of the team composition rule that a team must contain at least 14 Indian players. See appendix 1 for the team composition rules.

To avoid the problem of multicollinearity, principal component analysis was done on the input variables. Principal component analysis converts highly correlated variables into a set of linearly uncorrelated variables called principal components. The number of principal components created is less than or equal to the number of original variables. The first principal component captures the largest variance in the data followed by the second principal component and so on. To make our models simpler, fewer than all of the principal components were used as inputs as shown in figure 2 and figure 3. Beyond the selected PC ID (yellow line) in the figures, Eigen values are decreasing rapidly that show that the principal components with more than PC ID 12 in batting and PC ID 10 in bowling were not able to capture much variance in the data. Twelve principal components were thus selected for the batting dataset that captured 83% variance and ten principal components selected in the bowling dataset that captured 80% variance in the data.

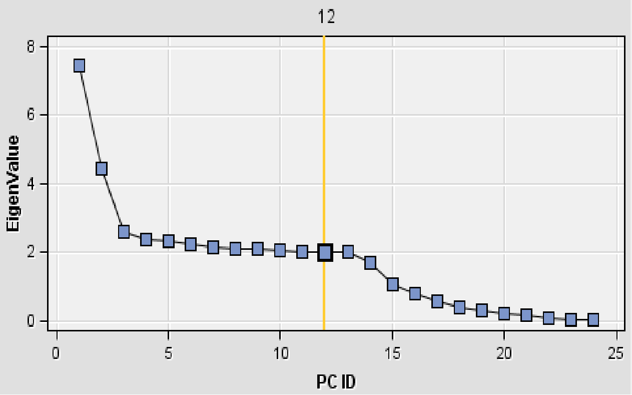

Figure 2. Batting PCA

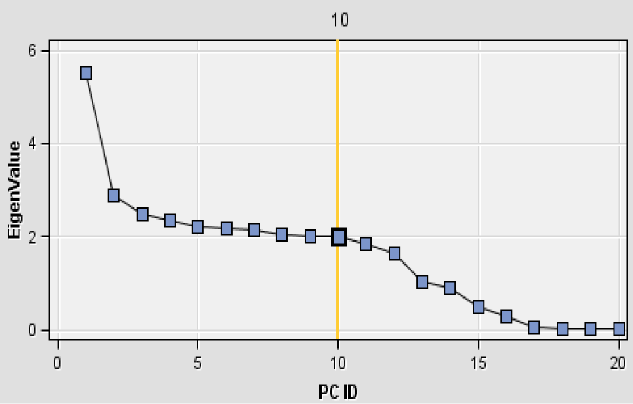

Figure 3. Bowling PCA

Output variables from the principal components analysis were used as inputs to the classification models. Various decision tree models, regression models and neural network models were constructed to develop the best model that could predict the selection of a player in the team. Various decision tree models with different parameters were developed, for example, different combinations of maximum branches, maximum depth, nominal criterion (Chi-square, entropy or Gini) and significance level for split were used. Logistic regression models were constructed with different combinations of polynomial degree and model selection method (Forward, backward or stepwise selection). Neural network models were built using combination of different architecture (Multilayer perceptron, ordinal radial or normalized radial) and number of hidden units.

Models developed were compared based on the validation misclassification rate. Finally the model that performed best was selected in each study and scored against the data containing the actual results of the selection of players in the 2012 season. The selected model in each study provided the probability of selection of each player that could be used as a valuation factor for a franchisee to make decisions for setting price of each player as described in next section.

BIDDING FUNCTION

There are various linear and non-linear bid functions (Gavious, Moldovanu, and Sela, 2002) that can be used by franchise for bidding in IPL for buying the players. Consider n number of owners bidding for an individual player. v_i is bidder i’s valuation for the player (prediction probability from selected predictive model), which is private from all other bidders. All bidders other than i perceive vi as a random selection out of the interval [0, 1], governed by the distribution function F, and independent of other valuations. We assume that F is continuously differentiable, and we denote by f the associated density function. We also assume that f(v)>0 for all v $\in$ [0,1].

A bid x causes a cost g(x), where g: R+→ R+ is a strictly increasing function. The bidder with the highest bid wins the player. We assume that the cost functions are linear, g(x) = x. Let there be n bidders face linear cost functions, then the bid function of every bidder is given by

$$b(v) = v\, F^{n-1}(v) - \int_0^{v} F^{n-1}(y)\, dy$$

$$0 \leq v \leq 1$$

RESULTS

Some of the important findings from exploratory analysis in both studies are presented below.

Main findings in the batting data:

1. Few players who have played more than 120 matches are all selected indicating all experienced players are included in the squads.
2. This form of cricket encourages youngsters; (A minimum of 6 players from the BCCI under-22 pool in each squad) thus, most of the selected batsmen have played less than 30 matches.
3. More than 70% of selected players have 50 or more as their highest score.
4. As per the rules for team composition (Minimum of 14 Indian players must be included in each squad), maximum players should belong to India and thus, most of the selected players are Indians.
5. Batting average is also an important variable and thus most of the players selected have an average more than 25.

Main findings in the bowling data:

1. Most of the Bowlers selected have a best bowling figure of 3 wickets per game or more.
2. Economy rate is a very important factor in all forms of cricket and here, most of the selected players have economy rate between 6.6 and 7.8, which is quite high.
3. More than 90% of selected bowlers played in the last season (2011).
4. Most of the selected bowlers have strike rate of around 20.
5. Most of the selected bowlers have an average between 18 and 30.

Franchises bid on a player, hoping to buy him with minimum possible spending. The selection takes place from a pool of players consisted of batsmen, bowlers and all-rounders. For the prediction of players, several classification models were developed as explained in the methodology section. 80% of the data was used for training the models and 20% for the validation. First of all modeling was done using raw variables excluding the highly correlated variable named Country. Some of the important variables found by the top models are presented in the table 4. It can be observed that 6s (Sixes) and Average were two most important variables in batting studies. Innings and SR (Strike rate) were mostly used by the models in the bowling studies.

Model	**Important Variables in Batting**	**Important Variables in Bowling**
Stepwise polynomial regression	Average*6s	Innings*SR and Average*Runs
Stepwise logistic regression	6s, Half Century	Innings
Entropy decision tree	Runs, Not out, 6s	Innings, SR, 4w
Chi-square tree	Average, 6s	Innings

Table 4. Variable Importance

Due to the high correlation of many input variables, principal component analysis transformation was performed in both studies. There was a substantial improvement in the results when new variables (principal components) instead of raw variables were used as input to the models.

Table 5 shows the results of the top 5 classification models developed for the batting study with raw variables and with principal components. There is a huge difference in the results of two modeling techniques. For the batting, the entropy decision tree model resulted in maximum overall correct rate of 87.3% in validation data, which shows that it was a high performance classification model. Splitting rules of the entropy decision tree were based on PC_3 being most important followed by PC_1, PC_7 and finally PC_4. It is interesting that the multi-perceptron (MLP) neural network model with three hidden units performed best among all in training data but was not able to give the best results in the validation data. It could not perform well in the validation data perhaps because it was over-trained in the training data. The same results can be seen in the ROC curve presented in the figure 4 where the neural network curve lays on the top of all for most of the time in the training but the entropy decision tree model wins the race during the validation.

Model	**Validation Misclassification Rate with Raw Variables**	**Validation Misclassification Rate with PCA**
Entropy decision tree	0.429	0.127
Neural network	0.524	0.174
Stepwise polynomial regression	0.460	0.190
Chi-square tree	0.429	0.206
Stepwise logistic regression	0.460	0.206

Table 5. Batting Modeling Result

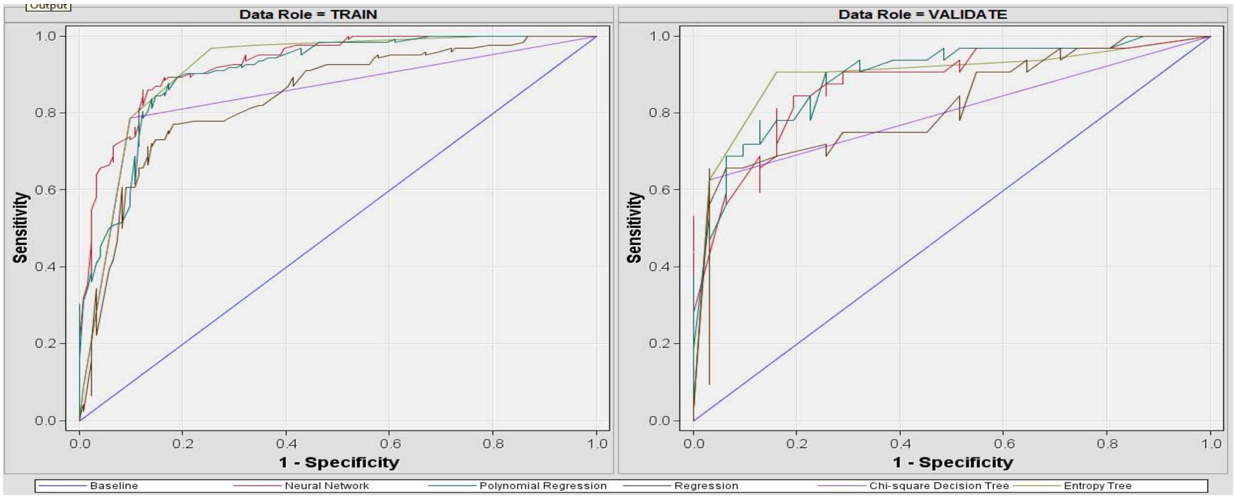

Figure 4. Batting ROC Curve

Table 6 shows the results of the models built using raw variables and principal components for the bowling dataset. The models constructed with the principal components performed much better than the models with the raw variables. The stepwise polynomial regression model with degree 2 including the main effects outperformed the classification models in the validation data by predicting the target variable 81.8% correctly as described in table 6. This model was built using 5 principal components (PC_1, PC_2, PC_1*PC_2, PC_2*PC_2, PC_2*PC_3). Again, it was noticed that neural network worked best among all for training data but could not do well with validation because of over training. It is on the top of all in the ROC curve in the training data but this time stepwise polynomial regression model did well in validation data based on misclassification rate as shown in figure 5.

Model	**Validation Misclassification Rate with Raw Variables**	**Validation Misclassification Rate with PCA**
Stepwise polynomial regression	0.473	0.182
Stepwise logistic regression	0.455	0.200
Entropy decision tree	0.455	0.200
Chi-square tree	0.473	0.273
Neural network	0.509	0.367

Table 6. Bowling Modeling Result

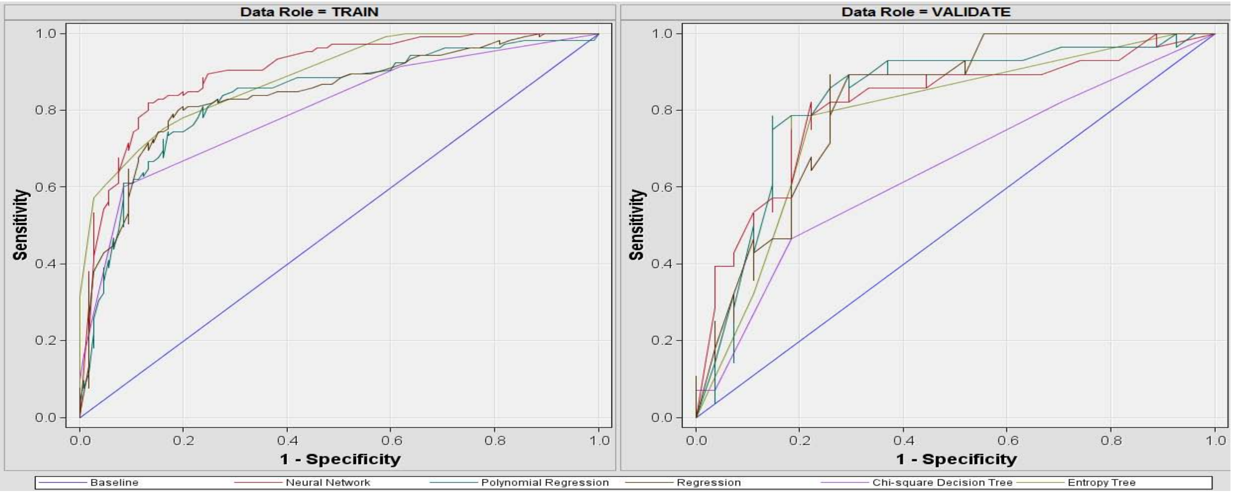

Figure 5. Bowling ROC curve

The model selected based on minimum validation misclassification rate in each study was scored against their respective data containing the actual results. Table 7 shows the statistics of performance of the entropy decision tree model selected for predicting selection of the batsmen. There were a total of 159 players selected in IPL and our model was able to predict 152 players correctly out of 159 i.e. 95.6%, which is really very high. Low value of false positive rate or 1-Specificity i.e. 24.03% high overall correct rate of 85.9% indicates that the developed model is robust.

Result			
	1	**0**	**Total**
Prediction			
1	152	37	189
0	7	117	124
Total	159	154	313
	Sensitivity = 152/159 = 95.60%	Specificity = 117/154 = 75.97%	Overall correct rate= (152+117)/313 = 85.9%

Table 7. Batting Scoring Results

Table 8 presents the results of the scoring of bowling population with the selected stepwise polynomial regression model based on minimum validation misclassification rate. The sensitivity of the model is 76.55% i.e. our model was able to predict 76.55% selected bowlers correctly which is a reasonable number. 1-Specificity of this model is only 16.66% and overall correct rate of 79.8% that indicates the high-class performance of the model. Selected model in each study performed extraordinarily well that proves that the predictive models developed could be used to help the franchise in making efficient selection decisions.

Scoring the data with best model in each study provided us the probability of selection of each player that can be effectively used as valuation factor in the bidding function as described in the methodology section.

Result			
	1	**0**	**Total**
Prediction			
1	111	22	133
0	34	110	144
Total	145	132	277
	Sensitivity = 111/145 = 76.55%	Specificity = 110/132 = 83.33%	Overall correct rate= (111+110)/277 = 79.8%

Table 8. Bowling scoring results

DISCUSSION AND FUTURE WORK

The results show that classification models can perform reasonably well in predicting the selection of players based on their past performance in the future seasons of Indian Premier League. Several predictive models were developed for the selection of batsmen and bowlers and based on minimum validation misclassification rate; the top most model was selected in each study. It was found that use of principal components instead of raw variables makes our models much efficient and robust. For the batting and bowling studies, the neural network model performed best for the training data with maximum overall correct rate but perhaps it was over-trained and could not produce best results in the validation data. In the batting studies, the entropy decision tree model outperformed the other models by providing more than 87% overall correct rate in validation data. Finally it was able to predict the actual selection of the players in 2012 season 85.9%. For the bowling data, the stepwise polynomial regression model with degree 2 did a good job in predicting the selection in validation data with only 18.2% misclassification rate and predicted the actual selection in 2012 season 79.8% correctly. The performance of the models was quite high which indicates that these can help decision makers during auction. By using our models, decision makers can reduce their list of players and thus make efficient selection decisions. The models developed provide a probability measure of selection of each player in the team. We suggest using the probability measure as a valuation factor in the bidding equation to set salaries for the players, as described in methodology section. It is also highly recommended for the team selectors in other sports as well to use principal component analysis when highly correlated variables are present in the data.

This study is the first attempt to develop predictive models for the selection of players in Twenty20 form of cricket. Iyer and Sharda (2009) did a similar research and found that the neural network model can predict the selection of players in the One-day international form of the cricket with the overall correct rate of 70%. As compared to their model, our models are doing better with more than 79% overall correct rate but in our study, form of the cricket is Twenty20 that requires different skill set. Beyond the accuracy of in predicting the selection of players in the IPL, these types of models can also be used to predict the selection in other forms of crickets or even in other sports.

This study being an exploratory project has some shortcomings. We included all-rounders in the batting dataset but results could have become more efficient if different studies were performed separately on them. So, we can say that our models can better predict the selection of only batsmen and bowlers. Fielding is another important aspect of cricket that was not considered as input in our study. There are some team composition rules set by the Board of cricket council of India for the IPL that were also not taken into consideration. Using Country variable forcefully as an input variable was an attempt to follow a rule that minimum of 14 Indian players must be included in each squad but specific rules were not implemented. There may be some non-playing factors like age, bid cap and other parameters that are not directly related to the playing abilities of the players that can influence the selection of the players were also not included in the modeling. So, the models developed can aid team owners in an objective way to make final decisions.

Pay cannot be adequately explained by past performance alone, nor are pay levels justified by future performance (Dalmia, 2010). So, the strategy of using probability of selection of players as a valuation factor in the biding equation can save franchise a lot of spending on players who are poor performers. Our models have the ability to build a talented team with minimum cost. Team selectors can use our models to make better decision on predicting the performance of the players in future.

Future work includes considering several other factors in the analysis like age, the ability to lead the team (captaincy), money spent on each player in past season, bid caps and performance of the players in other forms of cricket. Considering all these factors and team composition rules may result in better models.

REFERENCES

Beaudoin, D. (2003). The best batsmen and bowlers in one-day cricket. Thesis. Canada: Laval University.

Dalmia, K. (2010). The Indian Premier League: Pay versus performance. Thesis. New York University.

Gavious, A., Moldovanu, B., & Sela, A. (2002). Bid costs and endogenous bid caps*. RAND Journal of Economics, 33(4),* 709-722.

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

Karnik, A. (2009). Valuing cricketers using hedonic price models. *Journal of Sports Economics,* 11(4) 456-469.

Lenten, L. J. A., Geerling, W., & Kónya, L. (2012). A hedonic model of player wage determination from the Indian Premier League auction: Further evidence. *Sport Management Review*, 15(3), 60-71.

Parker, D., Burns, P., & Natarajan, H. (2008, October). Player valuations in the Indian Premier League. *Frontier Economics.*

Sharda, R. & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications,* 30, 243-254.

Singh, S., Gupta, S., & Gupta, V. (2011). Dynamic bidding strategy for players auction in IPL. *International Journal of Sports Science and Engineering,* 05(01), 03-16.

Singh, S. (2011). Measuring the performance of teams in the Indian Premier League. *American Journal of Operations Research,* 01, 180-184.

Zimbalist, A., S. (2002). Competitive balance in sports leagues: An introduction. *Journal of Sports Economics, 3(2),* 111-121.

APPENDIX 1

Some of the Team composition rules set by BCCI are:

1. Minimum squad strength of 16 players plus one physio and a coach.
2. No more than 11 foreign players in the squad and maximum 4 foreign players should be in the playing eleven.
3. Minimum of 14 Indian players must be included in each squad.
4. A minimum of 6 players from the BCCI under-22 pool in each squad.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Pankush Kalgotra, Oklahoma State University, Stillwater OK, Email: pankush@okstate.edu

Pankush Kalgotra is a second year graduate student majoring in Management Information Systems at Oklahoma State University. He has two-year experience of using SAS® tools for Data Mining, Texting Mining, and Sentiment Analysis projects. He is a SAS® certified Predictive Modeler using SAS® Enterprise Miner 6.1. In December 2012, he received his SAS® and OSU Data Mining Certificate.

Ramesh Sharda, Oklahoma State University, Stillwater OK, Email: ramesh.sharda@okstate.edu

Dr. Ramesh Sharda is a Regents Professor of Management Science and Information Systems at Oklahoma State University. He is also the Director of the Institute for Research in Information Systems at OSU and the Director of the Executive PhD in Business Program. His research interests are quite wide, with the general theme being application of analytical techniques and information technologies for decision support. Besides funded projects and numerous publications in this broad theme, he is a co-author of textbooks in Decision Support and Business intelligence areas, and co-editor of several book series.

Goutam Chakraborty, Oklahoma State University, Stillwater OK, Email: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is a professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU business analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.