GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

PERFORMANCE OF PREDICTIVE MODELS FOR CLUB MATCH VICTORY IN THE INDIAN SUPER LEAGUE: COMPARISON BETWEEN BOOTSTRAP AGGREGATION AND TRADITIONAL LOGIT

Vishnu Kurup M K[*1], Sachin M Varriar[2] & Prashobhan Palakkeel[3]
[*1&2]MBA-MS Student, Department of Management, Amrita Vishwa Vidyapeetham, Bangalore
[3]Assistant Professor, Department of Management, Amrita Vishwa Vidyapeetham, Bangalore

ABSTRACT

The purpose of this study is to identify the factors that are influencing the outcome of Indian Super League (ISL). This study mainly evaluates how factors such as attendance, market value of foreign players, the overall market value and other selected variables affect the outcome of each game. By answering the research questions, we identified the factors determining performance of football clubs in Indian Super League and based upon the results try to develop a model to predict the match results. A combination of statistical and machine learning tools is used in this study. The study evaluates three seasons of ISL during 2013-2016. The results show that there is no significant role for the attributes like attendance and average market value of foreign players in predicting the game outcome. However, the overall market value of players shows some level of significance. With the existing draft system, there is no real competition brought about by market value. The built predictive model using bootstrap aggregation (Random Forest) has achieved fair predictive capabilities with variables like difference in market value of home and away team.

Keywords: *Moneyball; ISL; Draft;.*

I. INTRODUCTION

Sports economics originated from United States and had a very low rate of penetration in Europe and Asia. In recent times major European and the Asian sports leagues are being evaluated and being researched. Most of the research in this area looks into the team and the league performance, match attendance, revenue from home games, broadcasting rights, labor market, competitive balance and uncertainty of outcomes and trends and the forecast of outcomes. For many years' Indian sports lacked serious attention as a corporate event, but with the introduction of manifold set of sports leagues, it is now capturing lot of noteworthiness. Sleeping giant, India woke up from the siesta with the commencement of Indian Super League (ISL) on 21st October 2013. The vision is to put football, as a promising sport in India and let Indian football sky rocket in the international football arena. The evanesced stars were given a chance, a Sophie's Career choice either to talk about it in the media area or to get back to the green carpet once again. ISL has now become one of the most important and 4th most viewed football league in the country. Taking the ingenuity from IPL, lot of leagues evolved in recent times. All these leagues looked into the customer's interests and other commercial aspects at different degrees. Sports industry have a large potential influence on country's economy. Revenues from the sports events, sale of sports merchandise and utilities are a huge source for tax for the country moreover it creates job opportunities. The corporate sponsorships for sports were dictated by cricket earlier, but other sports are also in the route to raise the same. When there was a 300% hike in on ground sponsorship for Kabaddi, it was 92%, 53.5% and 32 % in football, marathons and tennis respectively. In 2015 through Pro Kabaddi League, Star India reaped a 45cr revenue [1].

II. PREMISES OF THE RESEARCH:

A. Background History of Football analytics

Economy of sports has been a major field of study since the 1950s [2] though it has been restricted to the American countries. In recent years the eastern world especially, Europe is also looking at sports economics and its contribution to the economy [9]. A study on sports contribution to economy and employment was done by

SportsEconAustria who identified that Germany followed by UK are the countries with greatest GDP contribution from sports and highest employment in sports [3].

Apart from the research in baseball in the USA and research on football in Europe, in the recent past Asian and African countries have started to look at sports seriously and have started research on regional sport leagues [4] [5]. One study in US on the National football league evaluated the development of employment in a local area by the introduction of an NFL team. The results however didn't show any significant impact on the economy [10]. There is only very little research done on Indian football and its contribution to the economy. While sports have been contributing 1 to 5 % of GDP in most of the countries, in India the contribution is much less and there are no comprehensive studies on these aspects [6]. Research has been done on the value of brand alliances and the brand value a player brings into the team and league [7]. Researchers have analyzed how player market values change over time and studies shows that performances over the seasons and their age have close association in the top 5 European football leagues. The market value has been found to be a function of itself(auto-regressive) to a large extend with performance measures, fouls and age^2 adding very little value to the market value [8].

B. *Formulation of Hypothesis*

In this paper, we try to analyze two main issues, i) The factors determining the victory in football matches in the Indian super league ii) Role of money ball in Indian football, whether the Indian football victories depend on the market values of the players on paper. Another important thing to check here is Indian Super League's status as a competitive market in which teams are paying to improve squads and competing against each other. The assumption here is that India with its draft system does not promote competition and most teams are more or less equal.

The study attempts to develop a predictive model, which can be used to get hints about who would win a match. It can be used as a recommendation tool to tune the line up before each match or before each season during the auction.

DATA AND METHODOLOGY The data was collected from transfermarkt.com. The data about all fixtures and outcomes was collected. Each fixture is considered and the net value of players in the line-up of each match is found. The important variables considered for the study are Number of Foreign players - Home team (Foreign Home), Number of Foreign players - Away team (Foreign Away), Average Market Value of Home team (AvgMV Home), Average Market Value of Foreign players of Home team (F Avg MV H), Average Market Value of Away team (AvgMV Away), Average Market Value of Foreigner players of Away team (F Avg MV A), Average age – Home (AvgAge Home), Average Age – Away (AvgAge Away), match outcome (Win/unbeaten). The difference between the market values Total (MVDiff) and Foreign players (FMVDiff) are computed for each match.

We use both traditional logit and bootstrap aggregation models for the prediction. The performance of the models is compared based on accuracy, sensitivity and specificity.

Logit regression is a regression model in which the dependent variable is categorical, in most cases dichotomous. In our model 1 and 2, we use logit regression to find the odds of winning vs not winning. In model 3 and 4 we try to model to find the odds of being unbeaten and losing. In traditional logit [13] the probability of the dependent variable (here Win/unbeaten) can be defined as

ȁ ā g □ □ □ □□ We estimated four models as follows

Model 1

logit(Win)= ȕ0 + ȕ1 AvgMV Home + ȕ2 F Avg MV H + ȕ3 FMVDiff + ȕ4 MVDiff + İ

Model 2

logit(Win)= ȕ0 + ȕ1 FMVDiff + ȕ2 MVDiff + İ

Model 3

logit(unbeaten)= ü0 + ü1 AvgMV Home + ü2 F Avg MV H + ü3 FMVDiff + ü4 MVDiff + İ

Model 4

logit(unbeaten) = ü0 + ü1 FMVDiff + ü2 MVDiff + İ

Decision trees are a popular method of data mining. Often ensembles of decision trees are used to do classification or regression. Following Breiman[11], A. Muralidharan *et.al* [12] and S.Jaysri *et.al* [13] for bootstrap aggregation we can compute the ensemble as

[illegible]=

[illegible] —

1

Where *f*m is the mth tree.

Table 1. Descriptive statistics of the match outcome

Home Win	37
Home Loss	22
Home Draw	24

Table 2. Descriptive statistics of the variable used

	mean	SD	min	max
Foreign Home	5.98	0.15	5.00	6.00
AvgAge Home	28.83	1.05	26.60	30.70
AvgMV Home	191469.88	59309.92	102000.00	380000.00
FAvg MV H	362489.29	105786.15	195000.00	652500.00
Foreign Away	5.94	0.24	5.00	6.00
AvgAge Away	28.98	1.10	25.60	31.60
AvgMV Awa	196265.06	63857.71	78000.00	387000.00

y				
F Avg MV A	384513.14	128100.99	135000.00	708750.00
FMVDiff	-22023.86	159947.97	-420750.00	330000.00
AgeDiff	-0.15	1.53	-3.80	3.40
MVDiff	-4795.18	90115.61	-207000.00	207000.00

This study tries to identify the relationship between the market values, average age of squad and how it translates to a victory or remain unbeaten in a match. We use logistic regression for establishing the causal relationship among the variables. We also use a random forest algorithm to see if the match outcome can be predicted, considering the relatively poor performance of the logistic model. We can also get the variable importance for each variable while building the random forest. We use these techniques to check the existence of any causality on the outcome of the match. We also check whether the market value and age can be considered as the important factors for forming a squad.

We evaluate the performance of the models for their predictive capability using accuracy, sensitivity and specificity where,

Accuracy : (TP+TN)/Total matches
Sensitivity : (TP/(TP + FN))
Specificity : (TN/(TN+FP))

TP : True Positives
TN : True Negatives
FP : False Positives
FN : False Negatives

III. RESULTS AND ANALYSIS

The results of the analysis show that Model 1 is significant at 10% confidence level. Other 3 models aren't statistically significant. Hence it is observed that those variables may not necessarily be good predictors. Model 1 is used as a predictive model for match outcome. The coefficients of logit models can be seen in table 3 and 4.

Table 3. Logistic regressions for Victory

	Model 1 [AIC: 120.36]		Model 2 [AIC:118.17]	
	Coeff	P-value	Coeff	P-Value
Intercept	-1.208	0.3009	-0.1738	0.443
FMVDiff	6.103e-6	0.0923 *	3.547e-6	0.190
MVDiff	-1.307e-5	0.0707 *	-6.133e-6	0.200
AvgMV Home	1.113e-5	0.1988		
F Avg MV H	-2.989e-6	0.4648		

Table 4. Logistic regressions for unbeaten

	Model 3 [AIC:110.97]		Model 4 [AIC:105.2]	
	Coeff	P-value	Coeff	P-Value
Intercept	0.799	0.566	0.9015	0.000264
FMVDiff	7.331e-7	0.840	3.063e-7	0.913632
MVDiff	-4.288e-6	0.563	-2.577e-6	0.605053
AvgMV Home	2.163e-6	0.816		
F Avg MV H	-3.817e-7	0.931		

We considered 75% of the match data for training and 25% for validation. In addition to this, two random forest models are used for making prediction for Win and Unbeaten conditions. The comparison between model 1, random forest 1 and random forest 2 are tabulated below in table 5. It can be noted that all three attributes accuracy, sensitivity and specificity have considerably increased from model 1 to random forest 1, both models to predict win. Random Forest 2, modeled on Unbeaten matches was seen to be able to predict win and draws much more accurately having a 74% accuracy.

Table 5. Comparison of Logit model, Random Forest

	Accuracy %	Sensitivity	Specificity
Model 1	0.44	0.5294	0.2500
Random Forest 1	0.5217	0.6667	0.4286
Random Forest 2	0.7391	0.14286	1.00000

IV. CONCLUSION

The analysis presented in this paper show the problem of having a draft system in the Indian Super League as it becomes difficult to make a clear distinction about the quality of the team, as in the case of many European league systems. This hence is not a very good method for decision making or predicting match outcome based on line-up. Random Forest is seen as a better method to predict the Win/Draw in a match. The current Indian super league is a fair system giving equal chance to all clubs to win because of the draft system. It is exactly for this reason that a traditional logistic model fails to predict accurately the outcomes of a match, however a slight better prediction can be achieved using bootstrap aggregation.

This research is limited by the size of the data available. Only 3 seasons since the inception of the Indian super league, with 8 teams playing in a round robin format is used for the study. This lack of data could be a possible reason for the high specificity, low sensitivity of random forest 2. The league is also in the phase of expansion, with new teams joining the competition. Fatigue and related form is another variable that has not been considered in the scope of this paper. Future research can focus on these short comings.

REFERENCES

1. https://economictimes.indiatimes.com/magazines/brand-equity/war-of-leagues-with-ipl-isl-is-india-emerging-as-a-sporting-nation/articleshow/54672726.cms
2. S. Dobson and J. Goddard, The Economics of Football. Cambridge: Cambridge University Press, 2011.
3. SpEA, http://ec.europa.eu/assets/eac/sport/library/studies/study-contribution-spors-economic-growth-final-rpt.pdf
4. Ochieng', F. W. (2017). Factors determining the performance of football clubs in the Kenya Football League (Thesis). Strathmore University. Retrieved from http://su-plus.strathmore.edu/handle/11071/5551
5. P. Rattanapian, J. Tingsabhat and V. Kanungsukkasem, "Factors influencing achievement of Regional League Division 2 football tournament management", Kasetsart Journal of Social Sciences, 2017.
6. Sooraj Aurora, https://thediplomat.com/2016/07/indias-growing-sports-industry/
7. Y. Yang, M. Shi and A. Goldfarb, "Estimating the Value of Brand Alliances in Professional Team Sports", Marketing Science, vol. 28, no. 6, pp. 1095-1111, 2009.
8. O. Müller, A. Simons and M. Weinmann, "Beyond crowd judgments: Data-driven estimation of market value in association football", European Journal of Operational Research, vol. 263, no. 2, pp. 611-624, 2017.
9. Dave Russell. 05 Sep 2016, Routledge Handbook of Football Studies Routledge. Accessed on: 22 November 2017
10. M. Islam, "Local Development Effect of Sports Facilities and Sports Teams", Journal of Sports Economics, p. 152700251773187, 2017
11. Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), pp.123-140.
12. A. Muralidharan, Sugumaran, V., Dr. Soman K. P., and Amarnath, M., "Fault diagnosis of helical gear box using variational mode decomposition and random forest algorithm", SDHM Structural Durability and Health Monitoring, vol. 10, pp. 55-80, 2015.
13. S. Jaysri, Priyadharshini, J., Subathra P., and Dr. (Col.) Kumar P. N., "Analysis and performance of collaborative filtering and classification algorithms", International Journal of Applied Engineering Research, vol. 10, pp. 24529-24540, 2015.

.