Sustained Growth of Football Teams with Academy Training

- Proposal of Shapley-based Measurement -

Seiji Matsuhashi
Faculty of Economics
Gakushuin University
Tokyo, Japan
e-mail: 22122004ATgakushuin.ac.jp

Yukari Shirota
Faculty of Economics
Gakushuin University
Tokyo, Japan
e-mail: yukari.shirotaATgakushuin.ac.jp

***Abstract*—In this paper, the dominant factor for sustainable growth in football teams is described. Using the data of Japan-League teams, we conducted machine learning based regression and its interpretation by Shapley values revealed the Academy development was significant. In the financially large-scaled strongest teams, the Academy development is important to sustain the high scores/ranking. In small or medium sized teams, Academy development is another approach for growth to the upper league under limited budget. To measure the Academy development level, Matsuhashi's Measure based on Shapley values is proposed and it is proven that Matsuhashi's Measure has the highest correlation with the top strongest teams' winning points.**

Keywords-football; academy training; Shapley values; SHAP; effective growing; Matsuhashi's Measure.

I. INTRODUCTION

This study is a regression-based corporate evaluation analysis, and the subject of the analysis is football teams in the Japan-League (hereafter J-League), which will soon celebrate its 30th anniversary since 1993. The main evaluation indicator for football clubs is how effectively they perform under a limited budget. This is a common goal to every professional sports organization.

Investing a large amount of money for a high performance is a straightforward and instant approach. The first objective of this paper, however, is to explore another approach by which a small/medium sized team can effectively grow with a limited budget. The first author, Matsuhashi, has been interested in data analysis of the J.-League for many years. He has found some cases in which utilizing young players from the academy were important to the league performance for the teams. These clubs had small budgets in early years. However, owing to the results of such young players, the club increased its ranking, promoted to J1 League, and gradually expanded its scale. Then, in order to measure the Academy development level, Matsuhashi's Measure was defined in our previous work [1].

The second objective of the paper is to explore sustainability of the already large-scaled teams. In general, it is difficult to maintain high-performance in a company. In J-League, it is difficult for the strongest teams to keep the top positions. The paper showed that one of the driving forces is Academy development. In the paper, we will prove that Matsuhashi's Measure is useful to measure the large-scaled football team's sustainability.

We use machine learning regression analysis and Shapley values for interpretation of the result. The next section describes the data we used. Section 3 describes the analysis method using Shapley values. Section 4 explains Matsuhashi's Measure (MM) and shows the high correlation between the MM values and sustained high performance large teams. In Section 5, discussion concerning Matsuhashi's Measure is conducted. Finally, we conclude this paper.

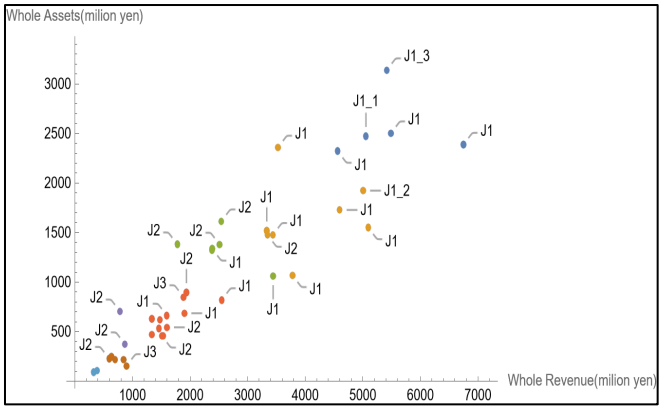

Figure 1. 10-year average operating revenues and total assets of the J-League clubs through 2021.

II. DATA AND METHOD

In this section, we explain the data used and the regression analysis method.

The cost of strengthening a professional soccer team is enormous in every country. The dominant part of the cost is the personnel cost for star players, but there is a disparity in financial scale of teams. Figure 1 shows a scatterplot between 10-year average operating revenues and total assets of the clubs from J-League. J1 is the top category, followed by J2 and J3. As shown in Figure 1, there is the tendency that the larger financial sized team, the higher ranking it has.

On the other hand, however, some medium-sized teams belong to J1. We found that they are committed to Academy development. Therefore, the following hypotheses were formulated for the growth pattern of small and medium-sized clubs.

 ISBN: 978-1-68558-056-8

Hypothesis: Smaller clubs with smaller budgets can take an approach to achieve higher performance by training young players in Academies.

We will conduct analysis to find such clubs. The analysis method is regression. The target variable is the annual ranking in the league converted into a score of 100 points, with 100 being the highest. As for the choice of explanatory variables, we examined the correlation coefficients among the managerial variables and found that many of them had multicollinearity, so we finally selected the following two explanatory variables.

Explanatory variables
(1) Salary costs: Personnel costs for the year.
(2) Academy operating costs: Total costs for the seven years prior to the target year (2021).

The impact of (1) salary costs' increase is instant but short. In contrast, (2) academy operation costs take long time until the effects appear. To see the effect, we use 7-year total cost of Academy operations. For example, when analyzing the 2021 season, the target value is the ranking scores in 2021and the salary cost data in 2021 are used. Concerning Academy operating costs, data from 2014 to 2020 are used. Data were taken from the site of J-League official [2]. This is about the financial data written in English, which includes all clubs belonging to J-League.

The regression analysis method was the XGBoost algorithm by scikit-learn package [3] [4]. Its GitHub site [5] has more information on the algorithm. Explanatory variables are, in advance, standardized for each variable shown in Figure 2.

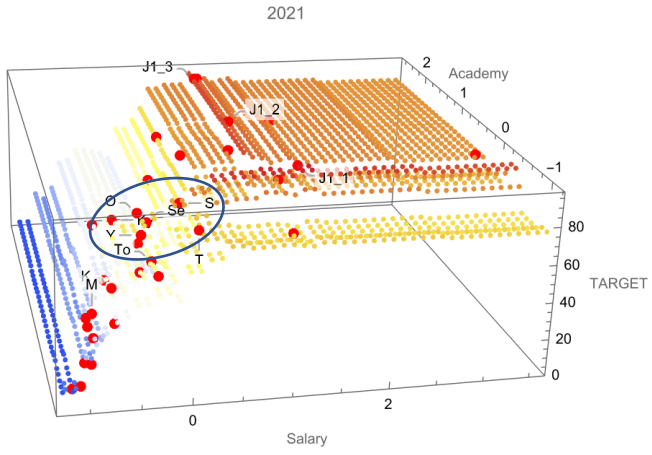

Figure 2. Resultant regression model with standardized salary and academy values and the ranking score as the target.

In Figure 2, the observation data represented by red points are plotted in a three-dimensional regression model *f(X)*. The warm colours such as brown correspond to higher target values.

In general, the more the salary and Academy operating costs, the higher the rank score. However, there are some exceptional medium-sized teams which have high rank scores even though their financial resources are limited. The teams will be later described in Section 4.

III. SHAPLEY VALUE

In this section, Shapley and SHAP values that we used are explained.

Shapley values are solutions in a game theory by multiple players [6-8] [9]. If n players work together, the profit (for example, 900 EURO) is divided to them. Then how they should divide the profit? How much is the individual player's contribution? Shapley found the unique solution of this question.

The main concept of Shapley values is **characteristic function** $\nu(X)$:

$$\nu : 2^n \rightarrow R$$

where n is the number of players and R means the profit by the subset of players. For example, there are three players A, B, and C. If A and B work together, the profit is 600 EURO. If C works together with A and B, then the profit becomes 900 EURO. The characteristic function defines the profit corresponding to any subset of players.

If n is 3, then the number of subsets is 2 x 2 x 2 = 8. In a real world (not in a theoretical world), it is too difficult to define the characteristic function. However, if such a characteristic function is given or can be found, each player's profit can be calculated using the following formula:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[\nu(x_{S \cup \{i\}}) - \nu(x_S)\right]$$

where ϕ_i is the Shapley value for player i, F is a set of players, and S is a subset of F which does not include i-th player, $S \subseteq F \setminus \{i\}$.

$|F|!$ is the permutations of the number of F. The term $[\nu(x_{S \cup \{i\}}) - \nu(x_S)]$ is player i's marginal contribution to the profit by S; the function ν is evaluated with the input of the player set $(x_{S \cup \{i\}})$ and then the function ν is evaluated with the input of the player set (x_S). The difference $[\nu(x_{S \cup \{i\}}) - \nu(x_S)]$ is the key part of the Shapley formula.

The term $|S|!\ (|F| - |S| - 1)!\ /|F|!$ expresses the appearance possibility of $x_{S \cup \{i\}}$. First S exists and then player i comes, and finally the left members of which number is $(|F| - |S| - 1)$ attend. Equality of appearance possibility is supposed here.

Finally, the sum of weighted differences is calculated that becomes the player i's profit. The easy explanation of the formula was described by Roth in [8].

Lundberg et al. modified the original Shapley value, so that we can use Shapley values in the machine learning regression analysis [10] [11] [12]. The customized Shapley value is called SHAP values. The differences are as follows:

(1) SHAP is defined for each explanatory variable i, instead of player i.
(2) Each data has its own **characteristic function**.
(3) The characteristic function is calculated using the regression prediction model $f(X)$

In this case of this paper, each football team has its characteristic function which is calculated using the regression prediction model $f(X)$. The regression model is visually shown in Figure 2. The vertical axis means the predicted target values.

Using the resultant regression model $f(X)$, an individual team's characteristic function $v_{team}(X)$ is calculated. **Function $v_{team}(X)$ predicts the team's target value for any subset of explanatory variables.** The characteristic function for each team reflects the team's behavioral characteristics.

If there is a missing value among the predictors, the value of *f(X)* cannot be calculated. Lundberg's idea is to input the average value of the predictor variable for the missing parameter [10]. By this idea, they could solve the problem that the characteristic function could not be defined in a real world. In Operational Management field, the concept of **an industry average value** is very important. In industry analysis, we firstly investigate whether a particular company's value is above or below the industry average. We think that this solution is reasonable from the industry analysis viewpoint as well.

In this paper, we express each predictor's SHAP value as "predictor_SHAP" such as **Academy_SHAP**.

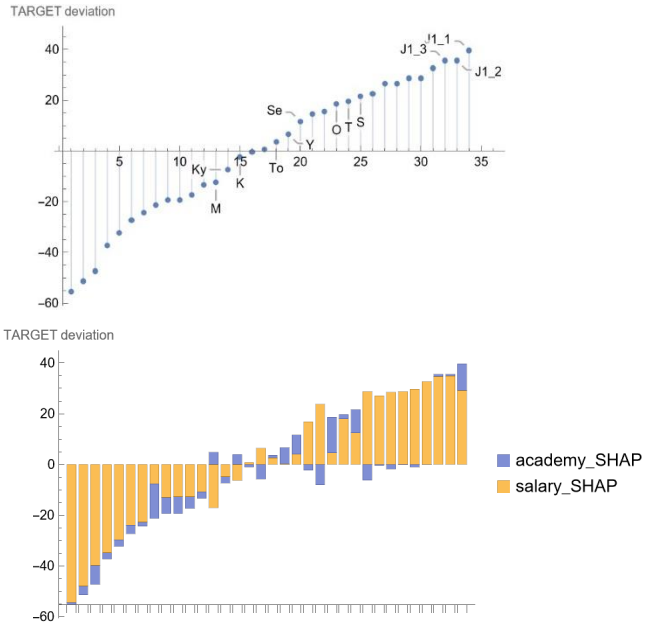

Figure 3. Target deviation values are divided to 2 SHAP values.

The resultant two SHAP values per team are visually illustrated in Figure 3. The x axis shows the 34 football teams. The y axis shows the deviation of target value which is its target value minus the average. The total SHAP values per team approximately becomes the deviation (see Figure 3).

Theoretically, the sum of SHAP values per data becomes equal to the target deviation of the data. However, if the fitting level of *f(X)* is low, the SHAP total is not equal to the deviation. In the lower bar chart in Figure 3, it is found that Salary_SHAP is much greater than Academy_SHAP. This means that in many teams the dominant factor of the target (ranking score) value is the salary expenses. However, some teams have significantly large Academy_SHAP. The ranking top team (see the right-end bar) has large Academy_SHAP, compared to others. In the next section, we shall evaluate the effect.

In a machine-learning based regression analysis, SHAP values are widely used for in various fields [13, 14]. Concerning football players, there are many researches using Shapley/SHAP values.

Sizov et al. use Shapley values to determine the salary prices or values of football players [15]. Hiller uses Shapley values to determine the performance/importance of players in each team of the Bundesliga in the season 2012/2013 [16]. Buzzacchi1 et al. use Shapley values to evaluate ranking of football managers in the Italian Serie A [17]. Marc Garnica-Caparrós uses SHAP values to understand gender differences in professional European football [18]. For example, it says that a high number of ground duels increases the probability of the model classifying a female player.

Although there are many researches by SHAP-based approach in the football field, as above mentioned, the target is the players' performance or managers' skills. As far as we know, there are **no** football team's management strategy evaluation by the SHAP approach. Our research is the first football teams' managerial structure evaluation by SHAP values.

In other fields except football, managerial researches by using SHAP values exist, as industry analysis [19][20] [21]-[23] [24].

IV. ACADEMY DEVELOPMENT LEBEL MEASUREMENT

In this section, we propose a measurement of academy training achievement level using the SHAP values described in the previous section.

First, we consider meanings of the Academy_SHAP value. Even if the Academy's operating expenses are large, if the Academy does not generate results, the ranking score does not increase and the Academy_SHAP value does not increase. In addition, even if players from the academy play more games, they do not always play important roles to obtain points of victory. Just Academy_SHAP is not sufficient to express the Academy development level, because there may be other hidden factors that contribute to the improvement in the ranking. In general, it is difficult to find causal relationships in policy and strategy evaluation [25].

To solve the problem, Matsuhashi defined the measurement for Academy development levels, based on the SHAP values. A new concept titled **"percentage of Academy graduates' participant ratio"** is introduced. This is the ratio of the number of appearances by players from the Academy to the number of league games available in a season (For example, for J1 in 2019, 34 games x 14 players). The calculation of the percentage of Academy graduates' participant ratio is defined as A/B where

- A: Total number of games from 2019 to 2021 played by Academy graduate players since 2011.
- B: The number of available slots for the three-year period from 2019 to 2021.

Matsuhashi's Measure = (Percentage of Academy graduates' participant ratio) x **(Academy_SHAP value).**

Next, we evaluate the performance of the measure.

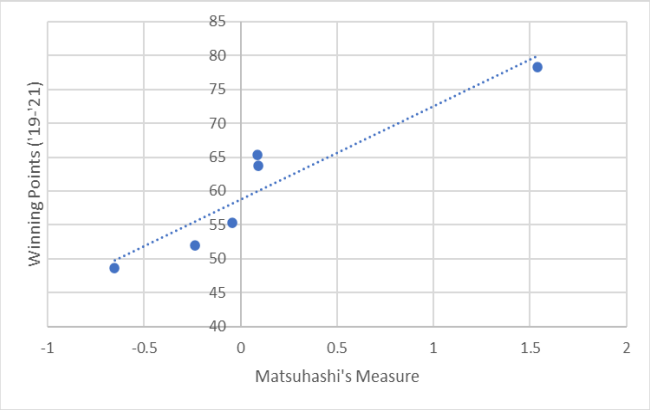

Figure 4. Relation between Matsuhashi's Measure and the average winning points with the correlation coefficient 0.94.

In Figure 4, the operating revenues top 6 teams in J-League are selected and the data are plotted. The x-axis shows Matsuhashi's Measure and the y-axis shows the average winning points from 2019 to 2021. These highest operating revenue teams are likely to be strongest ones as shown in Figure 1. The correlation coefficient of the relationship is 0.94 (see Figure 4). **This high correlation value indicates that the Matsuhashi's measure is highly correlated with the ranking score in the top group**.

The reason why the 3-year average winning points are used is that we would like to investigate the performance in the sustained period. From the relationship, we can say that **one of elements for maintaining strong teams is an excellent academy development base, and that Matsuhashi's Measure could measure the Academy development level with high accuracy**.

Reversely, what is the feature of the highest Matsuhashi's Measure team? In Table 1, the top 11 teams with the highest Matsuhashi's Measure values for three years through 2021 are listed. The names in yellow indicates a small or medium sized teams. In the scale-based clustering in Figure 1, these 11 teams are belonging to green or red clusters.

In the previous work, Matsuhashi described the followings [1]: *Although the ordinary strategy is to improve the ranking by increasing the financial investment, there were some medium-sized clubs that achieve high performance by investing in the Academy operation expenses. The Matsuhashi's Measure was effective in identifying these growing medium-sized clubs.*

TABLE 1. TEAMS WITH HIGHEST MATSUHASHI'S MEASURE VALUES.

Name	Matsuhashi's Measure ('19-'21)
S	2.074
J1_1	1.540
Y	0.842
O	0.677
T	0.323
Se	0.283
J1_2	0.091
J1_3	0.088
K	0.088
To	0.018
M	0.005

Using the resultant regression model, the highest Matsuhashi's Measure 11 teams are marked in the regression model (see Figure 5 and 6). The blue arrow marks depict the J1_1 to J1_3 which are the large-scaled teams in J1. The number of medium-sized teams in Table 1 is 8. These clubs have potential to be in a transition state to the large-scaled teams. Among them, the higher 5 teams are marked in the large circle, and the other lower 3 teams are marked in the small circle in Figure 5. The teams in the large circle can be identified ***growing medium-sized teams which produce high ranking scores.***

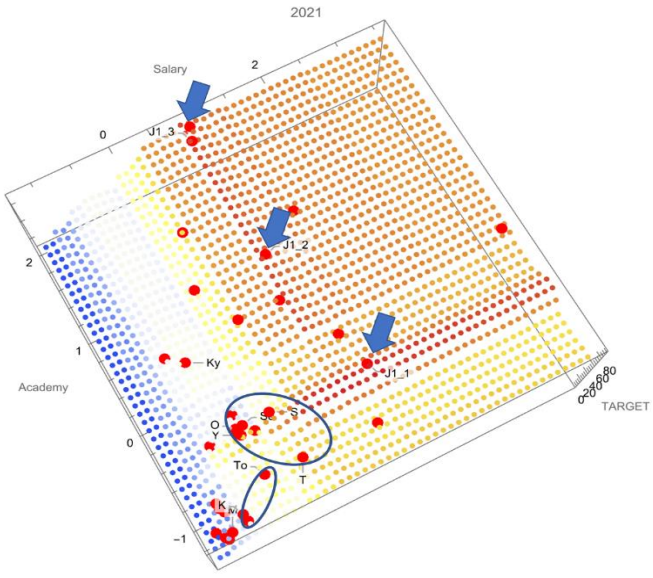

Figure 5. Highest Matsuhashi's Measure 11 teams on the regression model.

The closer look of the transition-state teams is shown in Figure 6. The 5 teams in the large circle have higher ranking score (target) values, compared with the teams in the small circle. Especially the team “S” with the highest Matsuhashi’s Measure has the highest target value. The medium-sized teams could be divided to the upper 5 teams and the lower 3 teams by target values as well as by the Matsihashi's Measure.

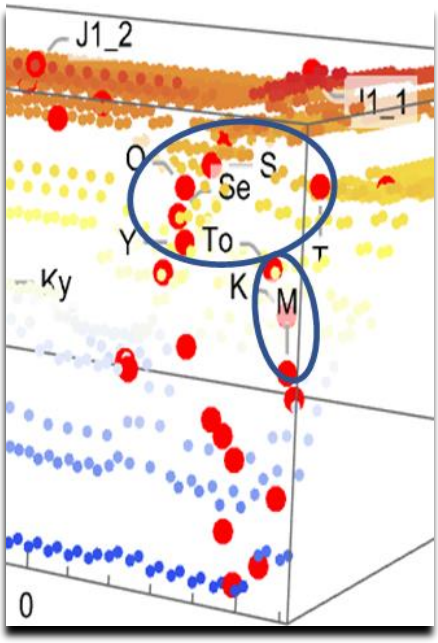

Figure 6. A closer look at highest Matsuhashi’s Measure teams on the regression model.

V. DISCUSSION

In this section, the reliability of Matsuhashi’s Measure (hereafter MM) will be discussed.

From the previous section result, it was found that the highest MM teams include

(1) the strongest three teams in 2021 which are large-scaled (J1_1 to J1_3), and
(2) effectively growing teams in 2021 which are medium-scaled.

When we compare MM values of J1_1 through J1_3, the strongest J1_1 team’s MM is larger by approximately 17 times (divide 1.54 by 0.091). One of driving forces of J1_1’s strength may be the Academy development level. Among the large-scaled teams, the high correlation 0.94 is found between MM values and the average winning points (see Figure 4). Among the large-scaled teams, Academy development level is important for the sustainability of the high performance and the level may be measured by MM.

Among medium-sized teams, there is high correlation between MM values and the target ranking scores (see Table 1 and Figure 6). For a medium-sized team, driving forces for growing is the Academy development level and the level may be measured by Matsuhashi’s Measure,

This study focuses on the relationship between academy development and annual ranking. The relationship represented by Matsuhashi's Measure may just a correlation and may not be a causal relationship. More analyses are necessary to prove the causal relationship [25]. This is our future research theme. At this stage, we can say that academy development is one of dominant factors to maintain its high performance for large-scaled teams and one of the effective approaches as a growth pattern for small and medium-sized teams.

VI. CONCLUSION

In this paper, we conducted a regression analysis of J-League results and analyzed the results using Shapley values, or more precisely, SHAP values. The novelty of the method is that the SHAP value is used to evaluate the contribution of the individual explanatory variables to the target value, taking into account the structural characteristics of individual football teams.

Based on the SHAP of Academy costs, the Academy development achievement measure named Matsuhashi’s Measure is defined as **(Percentage of Academy graduates’ participant ratio) x (academy_SHAP value).**

Among the large-scaled 6 teams, the correlation between Matsuhashi’s Measure values and average winning points was 0.94 which is very high. For the large-scaled teams, Academic development may be one of their sustainability factors. Then Matsuhashi’s Measure may measure the stability levels.

On the other hand, for medium-sized teams, Academy development gives another approach for growing. In such transition state teams with highest Matsuhashi’s Measure values, common feature can be found that they generate high performance effectively under the limited budgets. We can identify these teams’ positions visually on the regression model. These medium-sized teams have improved their scores by increasing the level of academy development achievement. This fact gives hope to smaller teams.

In conclusion, Academy development gives stability of high ranking to large and strongest teams and for the medium-sized teams, sustainable growth. In the paper, to measure the Academy development level, Matsuhashi’s Measure can be used with high accuracy.

Lastly, we shall describe the novelty of this analysis method. As company performance analysts, our interests exist on finding another approach to growth or success other than large investment. Observing the regression model, we may identify medium-sized but high-performance companies. Investigating the managerial states of these companies in the transition state, we would be able to find another approach strategy specific to the industry field. In the paper, the target industry field was the football team management. Then the recommended strategy was Academy training which was effective for sustained growth.

We shall continue to clarify the cause and result relation for high performance in football team strategies.

References

[1] S. Matsuhashi and Y. Shirota, "Finding of Corporate Growth Patterns by Shapley Values -- Case Study of Academy Development in the J-League -- " in *IEICE Technical Report* Kyoto, 2022/12/22-23 2022: IEICE, Technical Committee on

Infomation-Based Induction Sciences and Machine Learning (IBISML), p. (in printing).

[2] J. League. "Japan League Offcial Site. Individual Club Management Information | 公益社団法人 日本プロサッカーリーグ（Ｊリーグ） (jleague.jp) (accessed 2023/02/08).

[3] O. Kramer, "Scikit-learn," in *Machine learning for evolution strategies*: Springer, 2016, pp. 45-53.

[4] G. Hackeling, *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, 2017.

[5] XGBoostDevelopers. "XGBoost Decumentation (Revision 534c940a.)." https://xgboost.readthedocs.io/en/stable/ (accessed 2022/11/13.

[6] L. S. Shapley, "A value for n-person games, Contributions to the Theory of Games, 2, 307–317," ed: Princeton University Press, Princeton, NJ, USA, 1953.

[7] A. E. Roth, "Introduction to the Shapley value," *The Shapley value*, pp. 1-27, 1988.

[8] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[9] E. Winter, "The shapley value," *Handbook of game theory with economic applications*, vol. 3, pp. 2025-2054, 2002.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[11] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[12] S. M. Lundberg and S.-I. Lee, "Consistent feature attribution for tree ensembles," *arXiv preprint arXiv:1706.06060*, 2017.

[13] D. Lubo-Robles, D. Devegowda, V. Jayaram, H. Bedle, K. J. Marfurt, and M. J. Pranter, "Machine learning model interpretability using SHAP values: Application to a seismic facies classification task," in *SEG International Exposition and Annual Meeting*, 2020: OnePetro.

[14] A. Joseph, "Shapley regressions: A framework for statistical inference on machine learning models," presented at the King's Business School Working Paper, 2019.

[15] G. Sizov, P. Oztürk, and K. Valle, "The Use of Game Theory in Feature Selection."

[16] T. Hiller, "The importance of players in teams of the German Bundesliga in the season 2012/2013–a cooperative game theory approach," *Applied Economics Letters*, vol. 22, no. 4, pp. 324-329, 2015.

[17] L. Buzzacchi, F. Caviggioli, F. L. Milone, and D. Scotti, "Impact and Efficiency Ranking of Football Managers in the Italian Serie A: Sport and Financial Performance," *Journal of Sports Economics*, vol. 22, no. 7, pp. 744-776, 2021.

[18] M. Garnica-Caparrós and D. Memmert, "Understanding gender differences in professional European football through machine learning interpretability and match actions data," *Scientific Reports*, vol. 11, no. 1, pp. 1-14, 2021.

[19] K. Yamaguchi, "Feature Importance Analysis in Global Manufacturing Industry," *International Journal of Trade, Economics Finance*, vol. 13, no. 2, pp. 28-35, 2022. [Online]. Available: http://www.ijtef.com/vol13/719-UT0036.pdf.

[20] Y. Shirota, K. Kuno, and H. Yoshiura, "Time Series Analysis of SHAP Values by Automobile Manufacturers Recovery Rates," in *2022 6th International Conference on Deep Learning Technologies (ICDLT)*, 2022: ACM, pp. 135-141.

[21] K. Kuno and Y. Shirota, "Time Series Analysis of Shapley Values in Machine-Learning Regression," *IEICE Technical Report; IEICE Tech. Rep.*, 2022.

[22] T. Hashimoto, Y. Shirota, and B. Chakraborty, "SDGs India Index Analysis using SHAP," presented at the International Electronics Symposium (IES) 2022, Surabaya, Indonesia and online, 2022/8/9-11, 2022.

[23] M. Fujimaki, E. Tsujiura, and Y. Shirota, "Automobile Manufacturers Stock Price Recovery Analysisat COVID-19 Outbreak," in *POM Nara 2022*, Online, 2022.

[24] Y. Shirota, M. Fujimaki, E. Tsujiura, M. Morita, and J. A. D. Machuca, "A SHAP Value-Based Approach to Stock Price Evaluation of Manufacturing Companies," in *2021 4th International Conference on Artificial Intelligence for Industries (AI4I)*, 2021: IEEE, pp. 75-78.

[25] E. Duflo, R. Glennerster, and M. Kremer, "Using randomization in development economics research: A toolkit," *Handbook of development economics*, vol. 4, pp. 3895-3962, 2007.

 ISBN: 978-1-68558-056-8