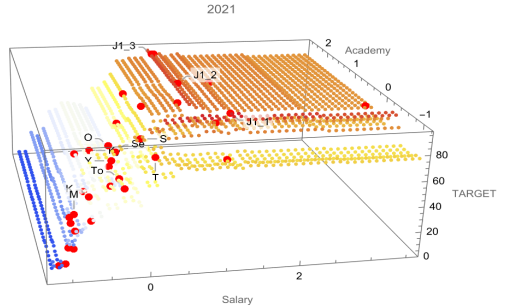
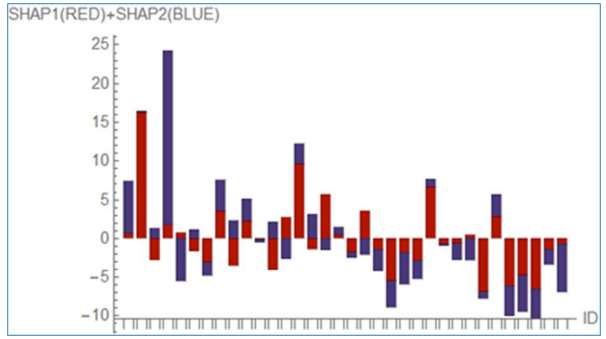
InfoSys 2023 Congress, March 13-17, 2023, Barcelona



TUTORIAL
Theoretical Explanation and Case Studies of Shapley Values in Machine Learning Regression



2023/3/13

- Prof. Yukari SHIROTA
  yukari.shirota@gakushuin.ac.jp
  Faculty of Economics, Gakushuin University (Japan)
- Prof. Basabi CHAKRABORTY
  basabi@iwate-pu.ac.jp
  Dean and Distinguished Professor, School of Computing, Madanapalle Institute of Technology and Science (India), and Prof. Emeritus of Iwate Prefectural University (Japan)

Biography

Prof. Shirota

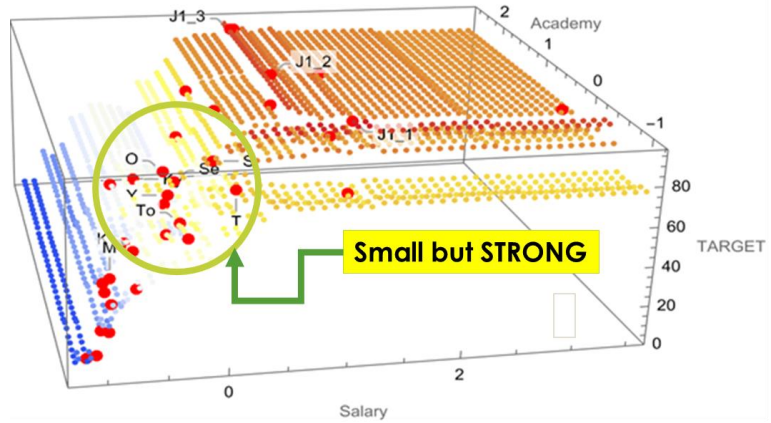
Professor of Gakushuin University. She graduated from the Department of Information Science, Faculty of Science, the University of Tokyo, and then received a D.Sc. in computer science in 1998. As a researcher in the private sector, she conducted research for 13 years and then in 2001 she was involved in Faculty of Economics, Gakushuin University, Tokyo as Associate Professor. In 2002, she has become Professor, Faculty of Economics, Gakushuin University. In 2006 to 2007, she stayed at University of Oxford, Oxford, UK as an academic visitor. She is Fellow of Information Processing Society of Japan. Research fields are industry analysis by AI, visualization of data on the web, social media analysis, and visual education methods for business mathematics. For over 23 years, she has developed visual teaching materials for business mathematics and statistics, and for mathematics used in AI (see the following sites):

- https://www-cc.gakushuin.ac.jp/~20010570/mathABC/SELECTED/
- https://www-cc.gakushuin.ac.jp/~20010570/SHIROTABASABI/
- https://www-cc.gakushuin.ac.jp/~20010570/mathABC/SELECTED/ShapeAnalysis/

Prof. Chakraborty

She received B.Tech, M.Tech and Ph. D degrees in RadioPhysics and Electronics from Calcutta University, India and worked in Indian Statistical Institute, Calcutta, India until 1990. She joined as a Faculty in the department of Software and Information Science, Iwate Prefectural University, Japan in 1998 and served as Professor and Head of Pattern Recognition and Machine Learning laboratory until her retirement in March, 2022. Currently she is a distinguished Professor and Professor Emeritus in Iwate Prefectural University. She also holds the position of Dean and Distinguished Professor in School of Computing, Madanapalle Institute of Technology and Science, Andhra Pradesh India. Her main research interests are in the area of Pattern Recognition, Machine Learning, Data Mining, Soft Computing, Text Mining and Time series analysis and their various real world applications in different fields such as Healthcare and Medical, Business and Finance, Social Media Data Analysis etc. She has authored more than 250 papers in reputed International Journals and peer reviewed International conferences. She is a senior life senior member of IEEE, member of ACM, Japanese Neural Network Society (JNNS), Japanese Society of Artificial Intelligence (JSAI). She is an active member of IEEE WIE (Women in Engineering) affinity group, chaired IEEE WIE Japan Council in 2010-2011 and founding chair of IEEE WIE Sendai in 2017-2018. Currently she is a member of IEEE R10 ARC and R10 SPNIC committee.

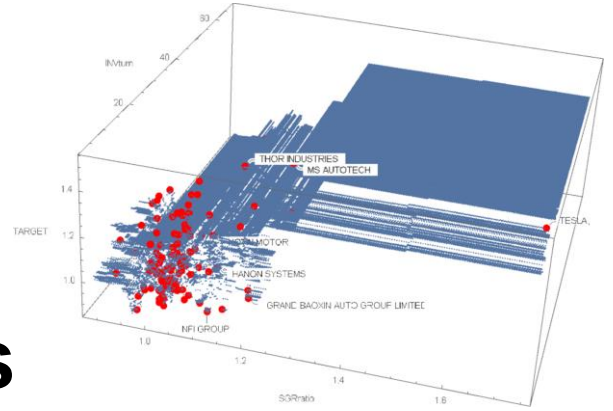
Topics of research interest of current projects:



- **Football teams' managerial evaluation: How to make a strong team even if currently a small team that is**
- **Method: AI-based regression plus Shapley values' evaluation**
- **Data: Currently only Japan-League's data**
- **Data request: Other country's team management data (Academy operation costs, NetSales) ,winning point data and players' appearances data**
  e.g. LaLiga Santander and Campeonato Nacional de Liga de Segunda División, and other countries' league data

To yukari.shirota@gakushuin.ac.jp
With subject "FOOTBALL"

Contents

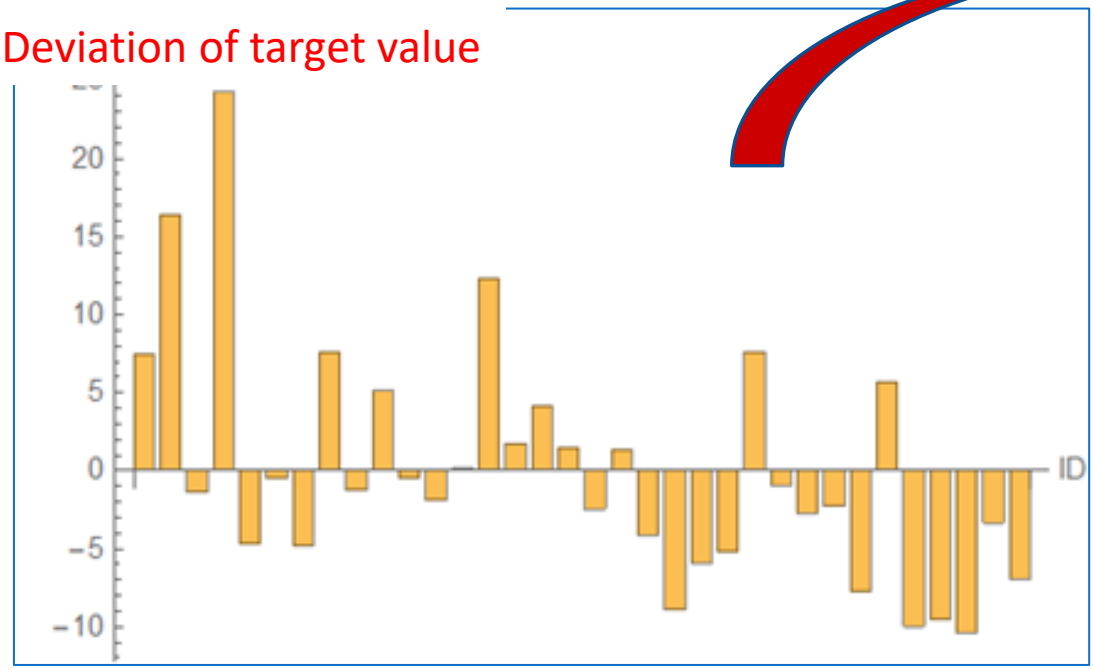


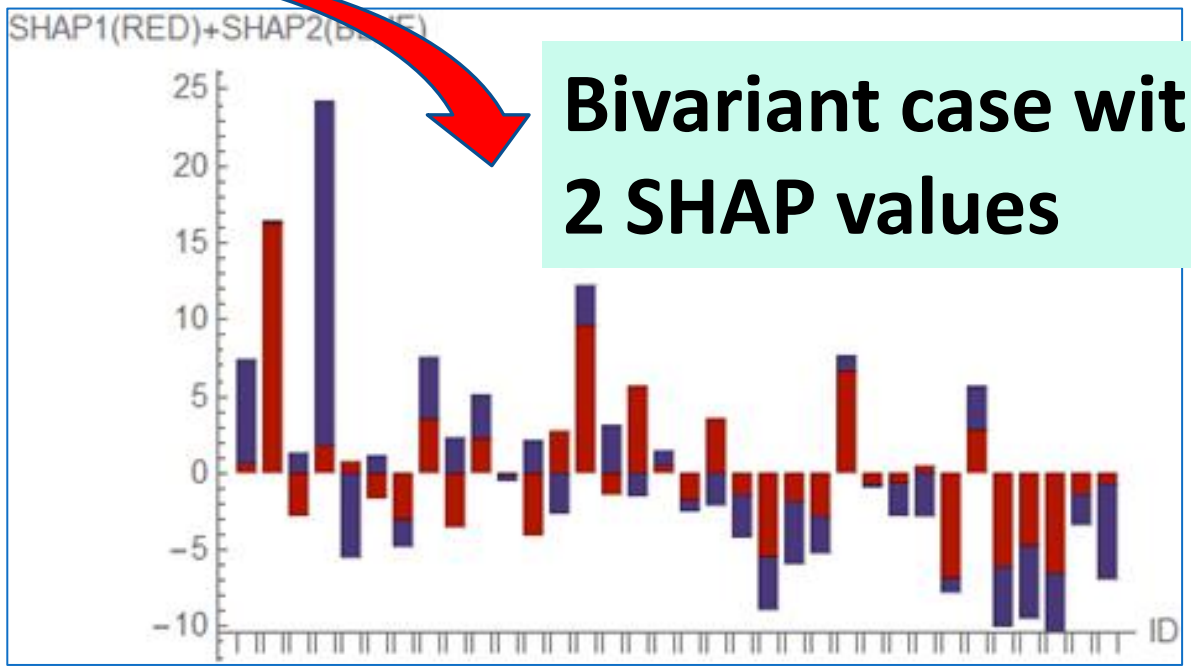
Shapley value evaluation after regression

Q:Which variable is the dominant factor？

Deviation of target value is divided to SHAP values

Deviation of target value

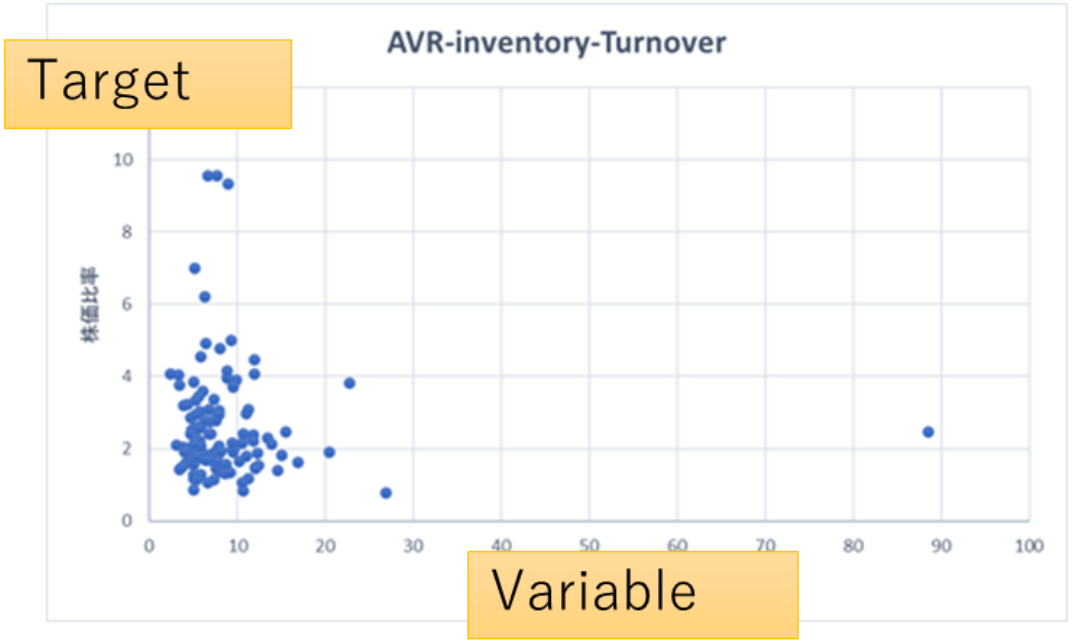




Bivariant case with 2 SHAP values

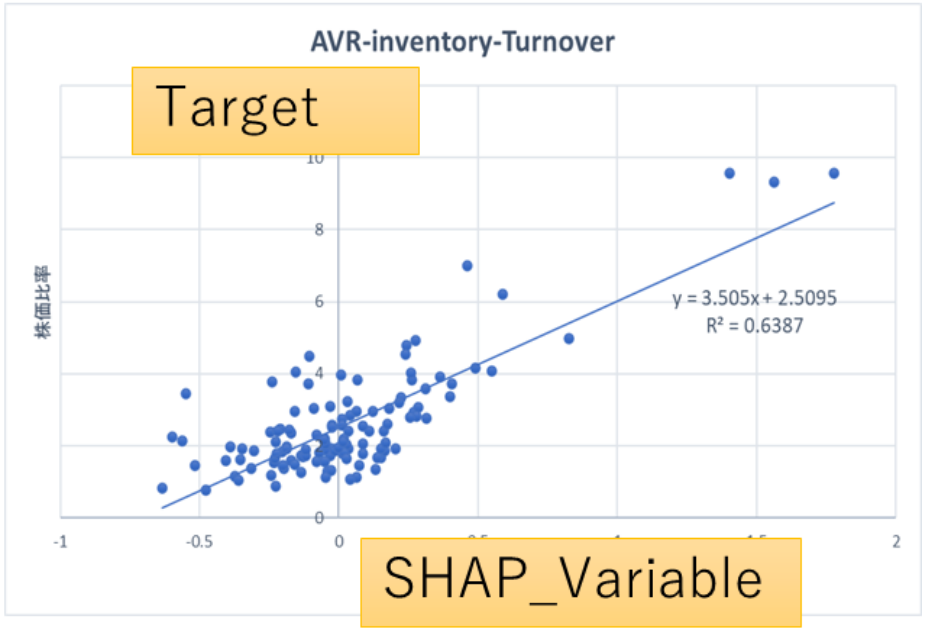
Correlation becomes higher if Shapley used

because characteristics are used



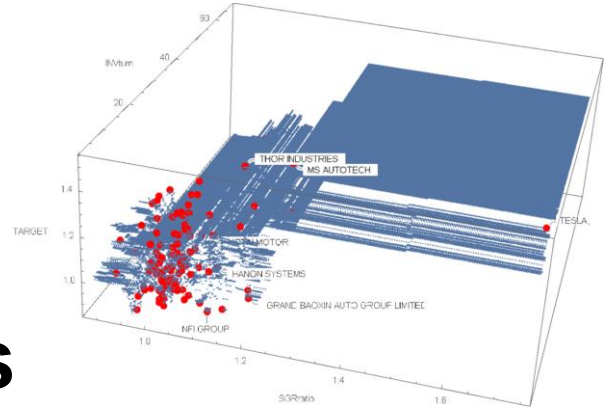
No relationship

Using SHAP of VARIABLE



High correlation coefficient

Contents



From original Shapley values to Lundberg's SHAP values

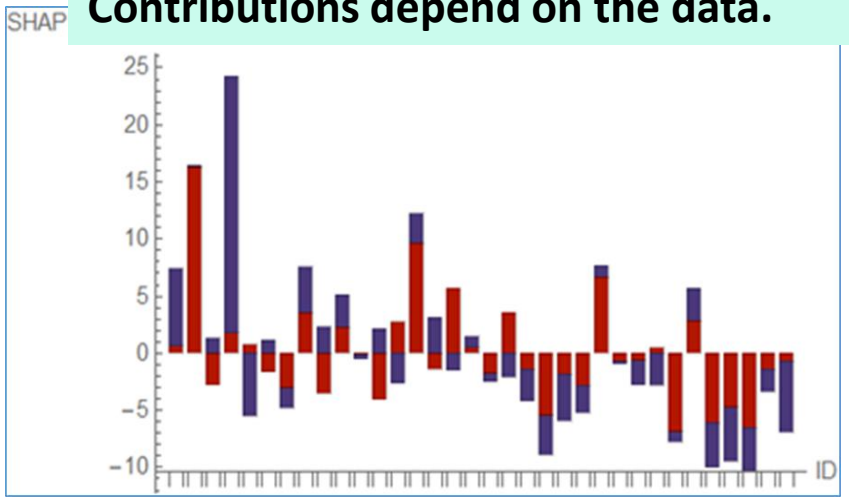
N players cooperative game

- Co-working by Lucía and Sofía, payment becomes $200
- By Lucía, Sofía, and María, payment becomes $500
- How to distribute the payment to members
- The unique solution of the game is Shapley's formula

N variables regression analysis

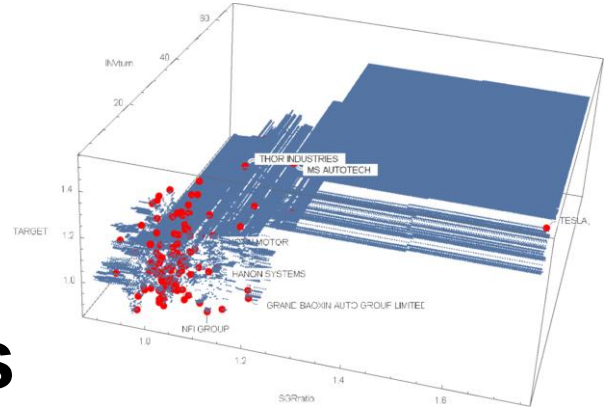
- A variable is a player
- In each data, the target value is the payment
- Each data has a different target value → For each data, SHAP values are calculated
- Ex. In 5th data, SHAP_variable #1 (or #2)

Contributions depend on the data.



Shapley values are positive but SHAP values may be negative

Contents



Characteristic Function ν

For subset of players (variables), it returns the payment

$$\nu: 2^n \rightarrow \mathrm{R}$$

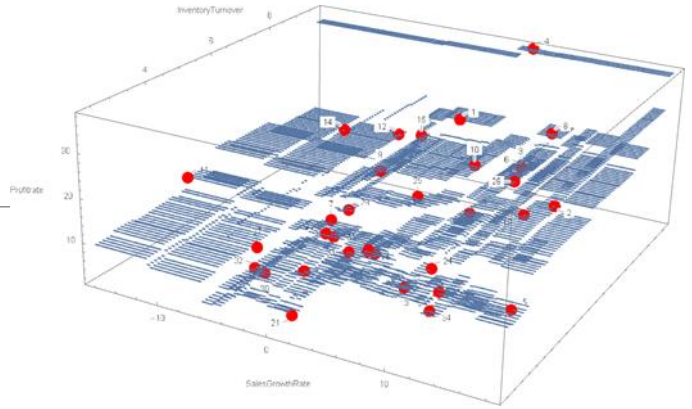
Given 5 players, 2^5 payment values are required.

1. SGR : Sales Growth Rate[%]
2. ROE[%]
3. ROA[%]
4. INV : Inventory Turnover Ratio[times/year]
5. FA : Fixed Asset Turnover Ratio [times/year]

But we cannot find a payment for any set of variables.

For a long time it was not possible to use Shapley values in real concrete problems

Lundberg solved the problem



- **Create pseudo-characteristic function by regression model f(x)**

If a missing parameter exists, *f(X)* cannot be calculated…

- **Expected (average) values for missing variables**

The concept of industry average is also important in the evaluation of companies. This approach is reasonable.

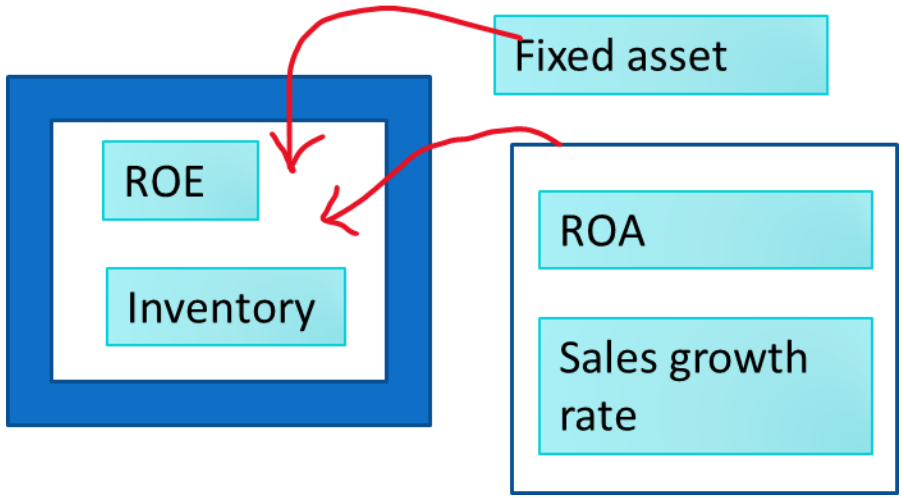
$\phi_{k,i}$ in k-th data, i-th variable's Shapley value

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\ \ (|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

Variable i's contribution after S

- Explanatory variables join one by one
- The whole permutations becomes $|N|!$ (factorial)
- Assumed to occur with equal probability

$$|S|! \times 1 \times (|N| - |S| - 1)!$$



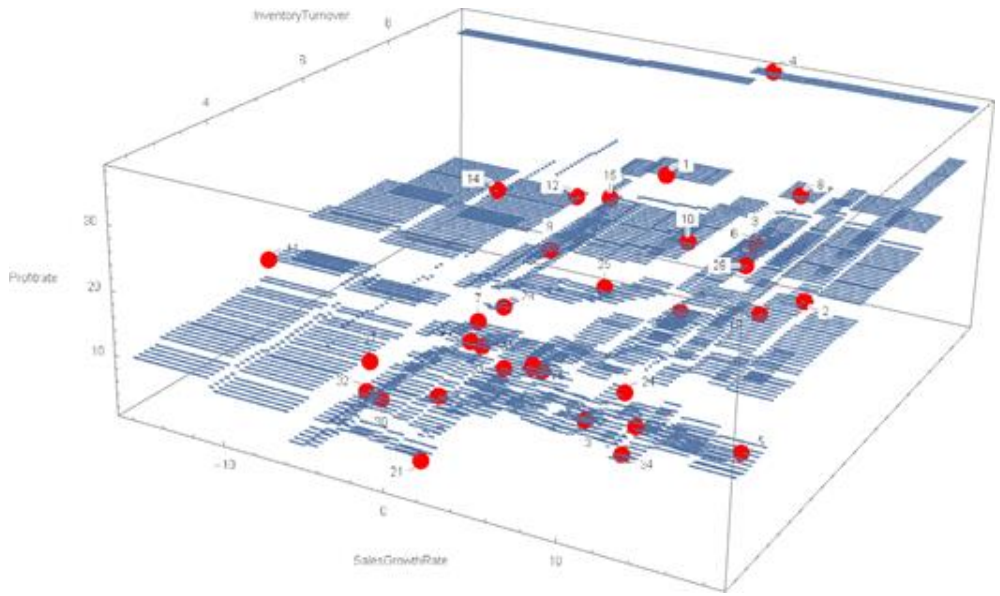
Expected (average) values for missing parameters
Bivariant regression model f(X)

To calculate v({var_#1})

f(var_#1, Average_#2)

To calculate v({}) (null set's payment)

f(Average_#1, Average_#2)



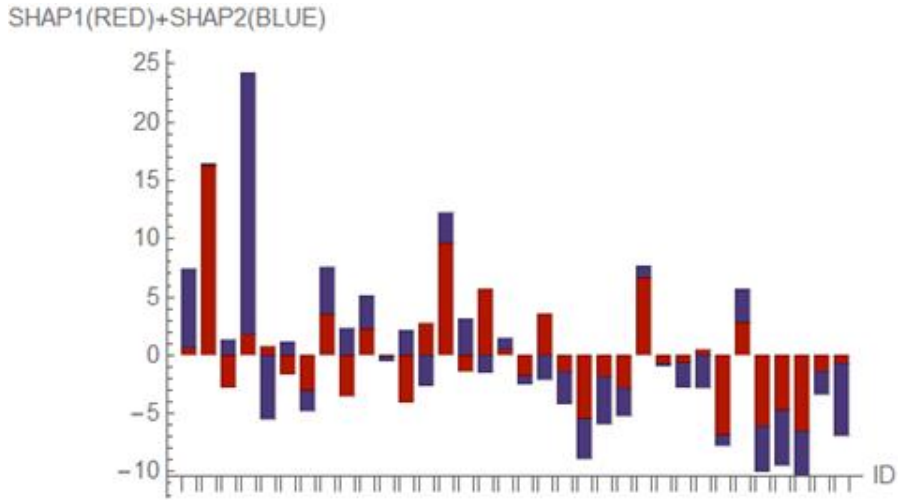
Expected (average) values for missing parameters
Bivariant regression model f(X)

Explanatory variables : SGR and INV(Inventory Turnover Ratio)

S is null set, S is just {INV}

$$\phi_{SGR} = \frac{1}{2}[f(SGR, 5.5) - f(2.9,\ 5.5)] + \frac{1}{2}[f(SGR, INV) - f(2.9, INV)]$$

$$\phi_{INV} = \frac{1}{2}[f(2.9, INV) - f(2.9,\ 5.5)] + \frac{1}{2}[f(SGR, INV) - f(SGR, 5.5)]$$

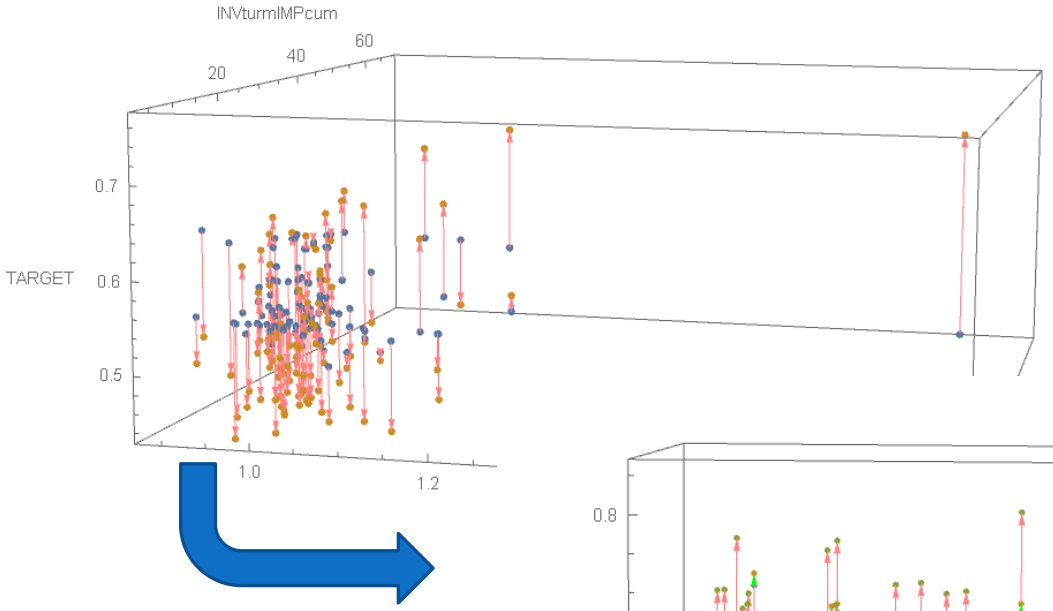


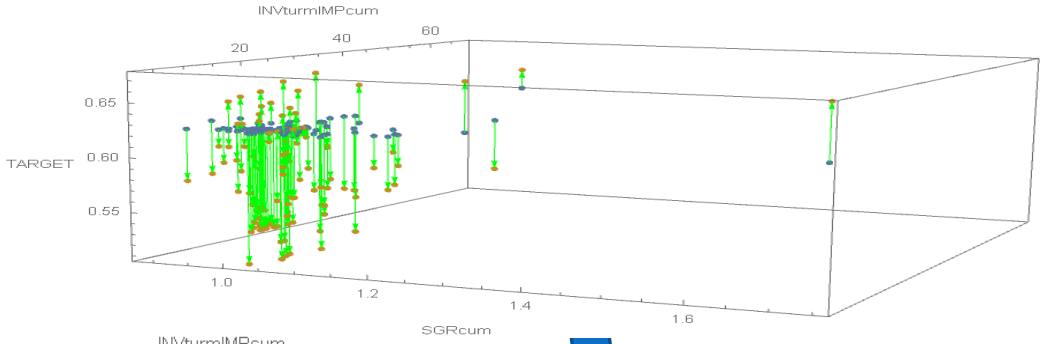
Bivariant regression model f(X) ϕ_{SGR}

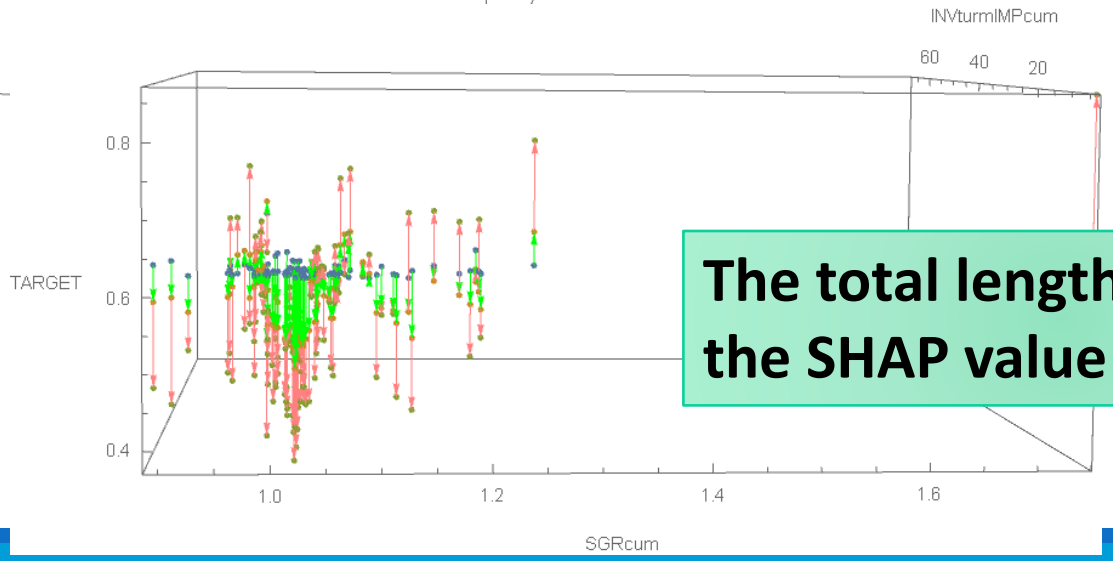
Explanatory variables : SGR and INV(Inventory Turnover Ratio)

$$\frac{1}{2}\left(f(SGRratio, INVturnover) - f(1.08, INVturnover)\right)$$

$$\frac{1}{2}\left(f(SGRratio, \quad 9.07) - f(1.08, \quad 9.07)\right)$$







The total length is the SHAP value of SGRatio.

SHAP based on the data characteristics

Using characteristic functions, each variable's contribution to target value is evaluated

The traditional regression evaluated only the total trend.

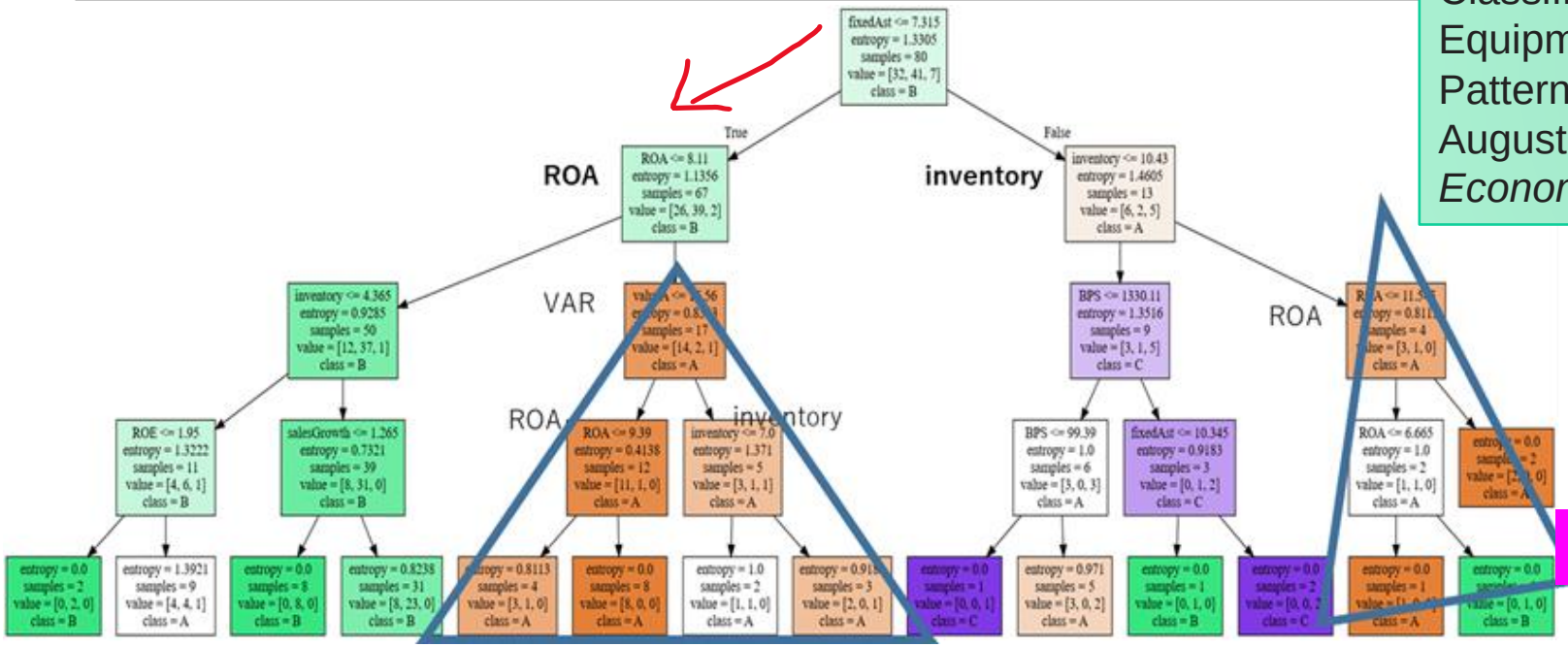
SHAP advantage: characteristic evaluation
-> Internal structure of the data
-> Better than judgment by absolute values

Example in a medical field:

- **The incidence of disease differs from person to person, even with the same sleep duration and dietary environment.**
- **Physical characteristics should be used**

Possible another approach for high performance

YAMAGUCHI, Kenji; SHIROTA, Yukari. Classification of Japanese Electrical Equipment Manufacturing Industry Recovery Patterns after Disasters: Case Study of August 2019. *International Journal of Trade, Economics and Finance*, 2020, 11.6.

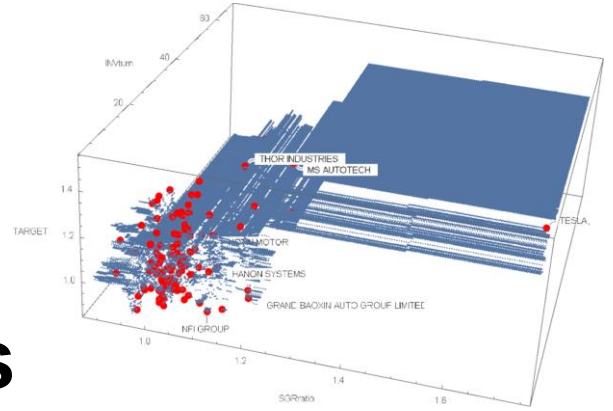


Normal approach to high class A

Another approach to high class A

Even if the tangible fixed asset turnover ratio is below 7.3, the probability of being Class A is high, if the ROA is above 8.1.

Contents



1. Graphical explanation of Shapley values
2. Cooperation game by explanatory variables
3. Theory of Shapley values
   A) Formula of Shapley values
   B) Case of bivariant
4. Case1: Time Series Analysis of SHAP Values by Automobile Manufacturers Recovery Rates
5. Case2: Football Teams Sustained Growing by Academy Training
   - Proposal of Shapley-based Measurement –
6. Conclusion

6th International Conference on Deep Learning Technologies (ICDLT 2022)

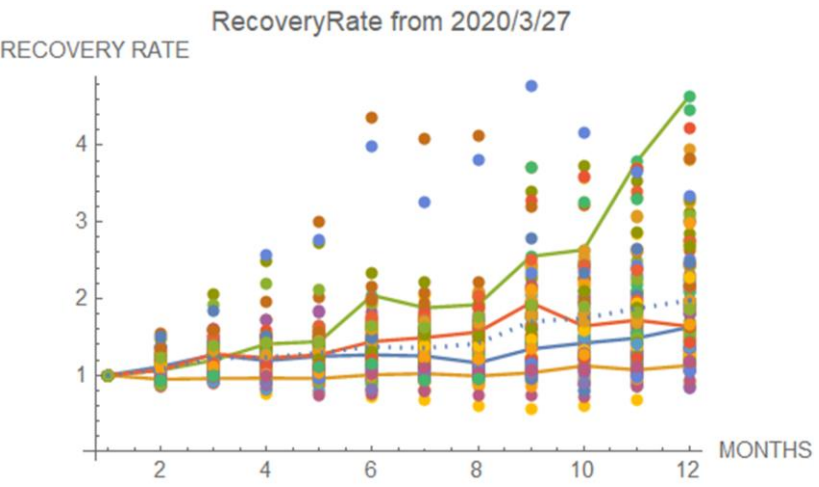
Time Series Analysis of SHAP Values by Automobile Manufacturers Recovery Rates

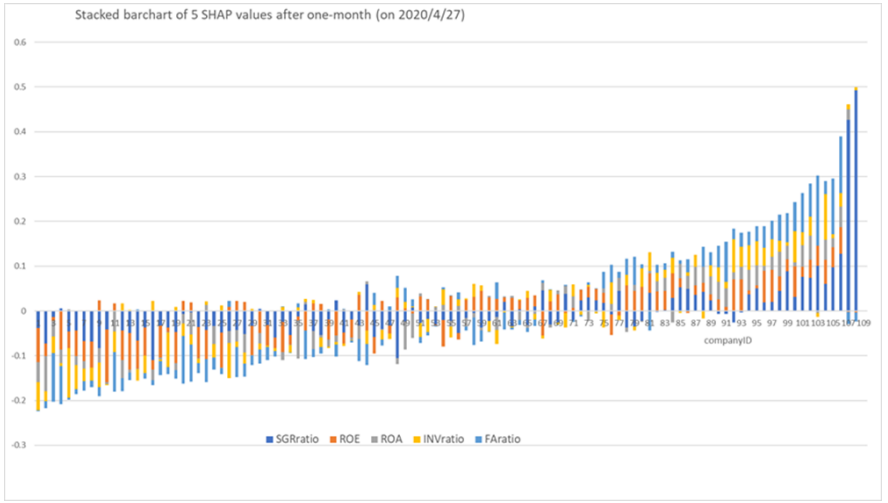
2022/07/22

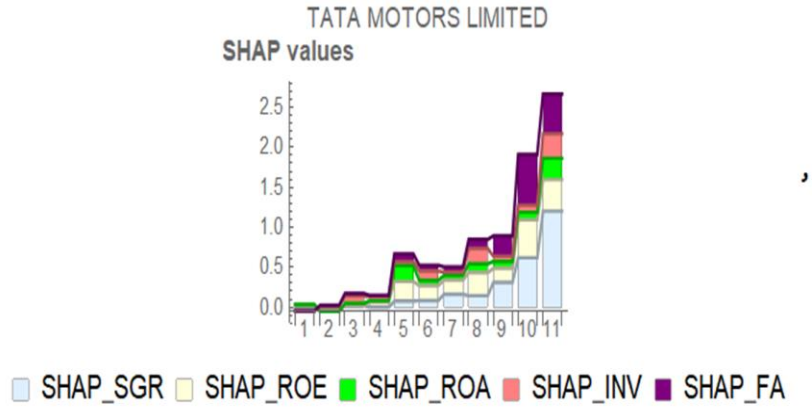
Prof Yukari SHIROTA (Gakushuin University)

Mr Kotaro KUNO (Gakushuin University)

Prof Hiroshi YOSHIURA (Kyoto Tachibana Univeristy)

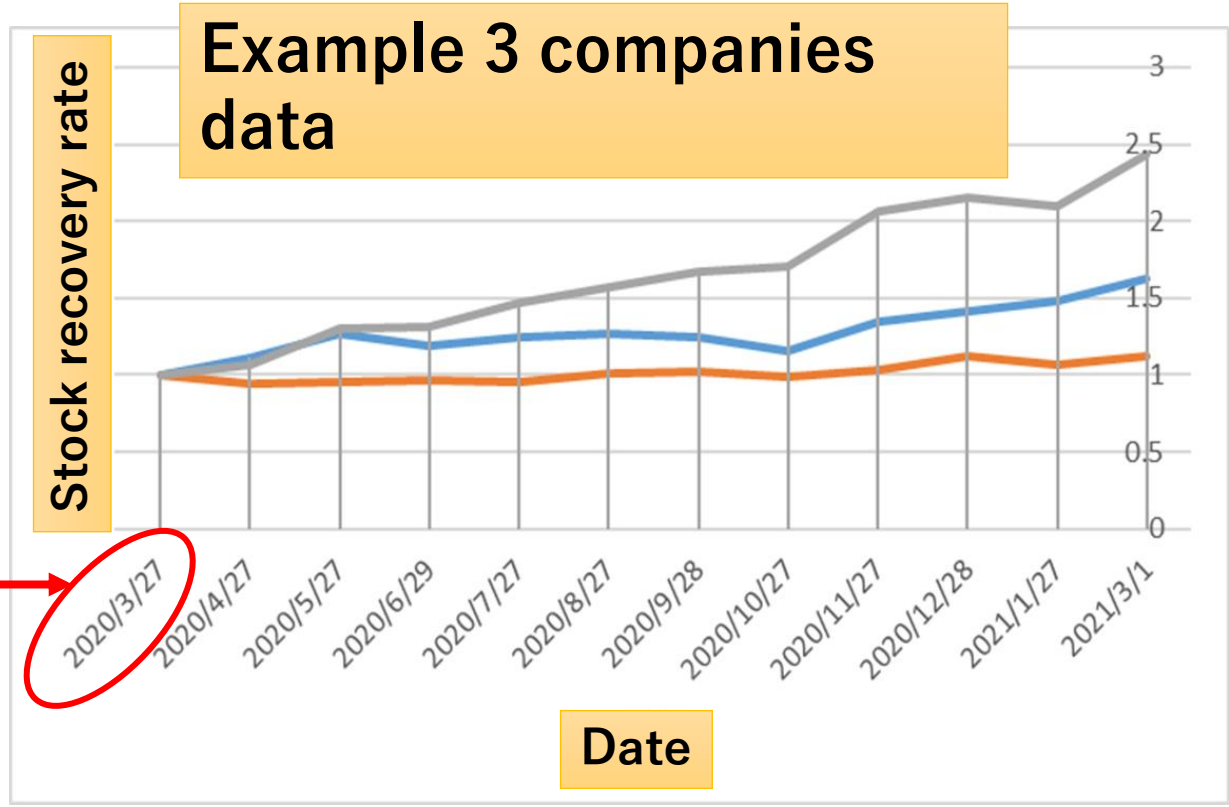






Research Objective

- 108 Global automakers stock price data at the outbreak of COVID-19
- From 2020/03/27 many companies started to recover stock prices
- What is important factors for the recovery ?
- Regression analysis
- Target variable: Stock recovery rate with the bottom value as 1



Regression

Target variable

StockRecoveryRate_i ＝

[Stock Price after (i) months from 2020/3/27] ÷ [Stock Price on 2020/3/27]

(i=1…11)

5 Predictor Managerial Variables from ORBIS DB

1. SGR : Sales Growth Rate[%]
2. ROE[%]
3. ROA[%]
4. INV : Inventory Turnover Ratio[times/year]
5. FA : Fixed Asset Turnover Ratio [times/year]

Predictor data set

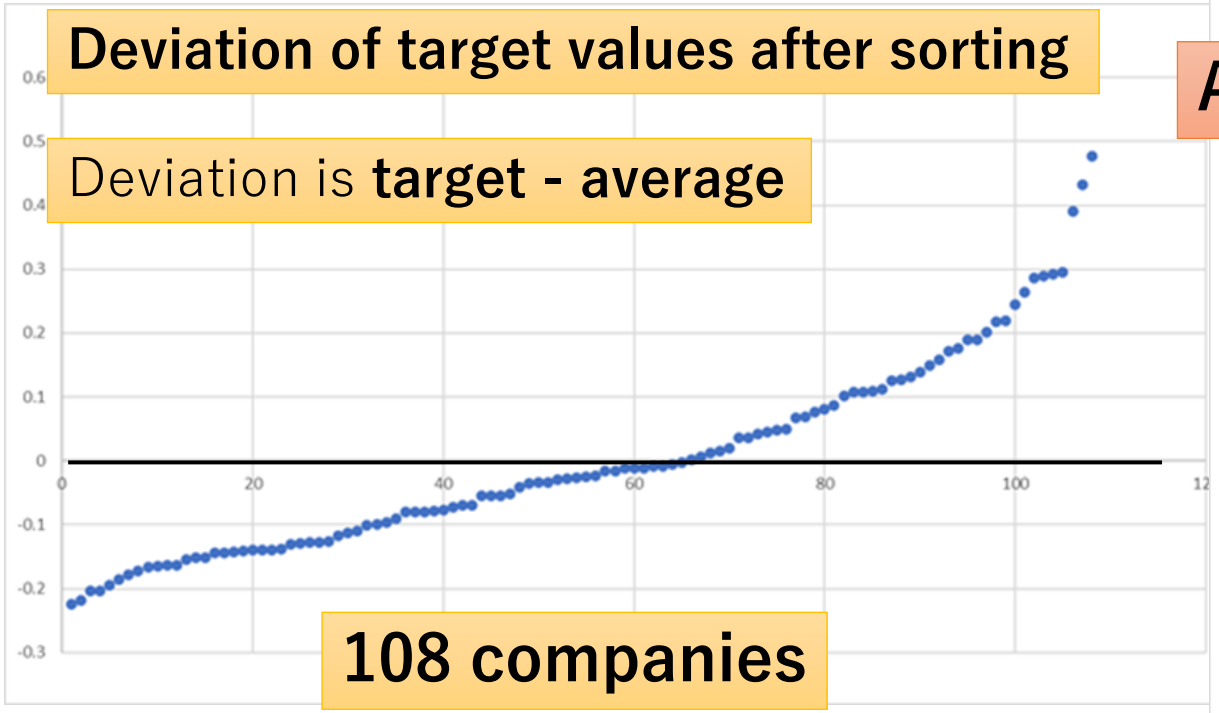
- **Average of 10 annual data**
- **Same data set is used for 11 regressions**
- **Suppose that companies' behavioral structures cannot be changed easily**
- **Long period at least a 5-year period is needed to ignore some events in a specific year.**

1. SGR : Sales Growth Rate[%]
2. ROE[%]
3. ROA[%]
4. INV : Inventory Turnover Ratio[times/year]
5. FA : Fixed Asset Turnover Ratio [times/year]

Deviation divided into SHAP values

Deviation of target values after sorting

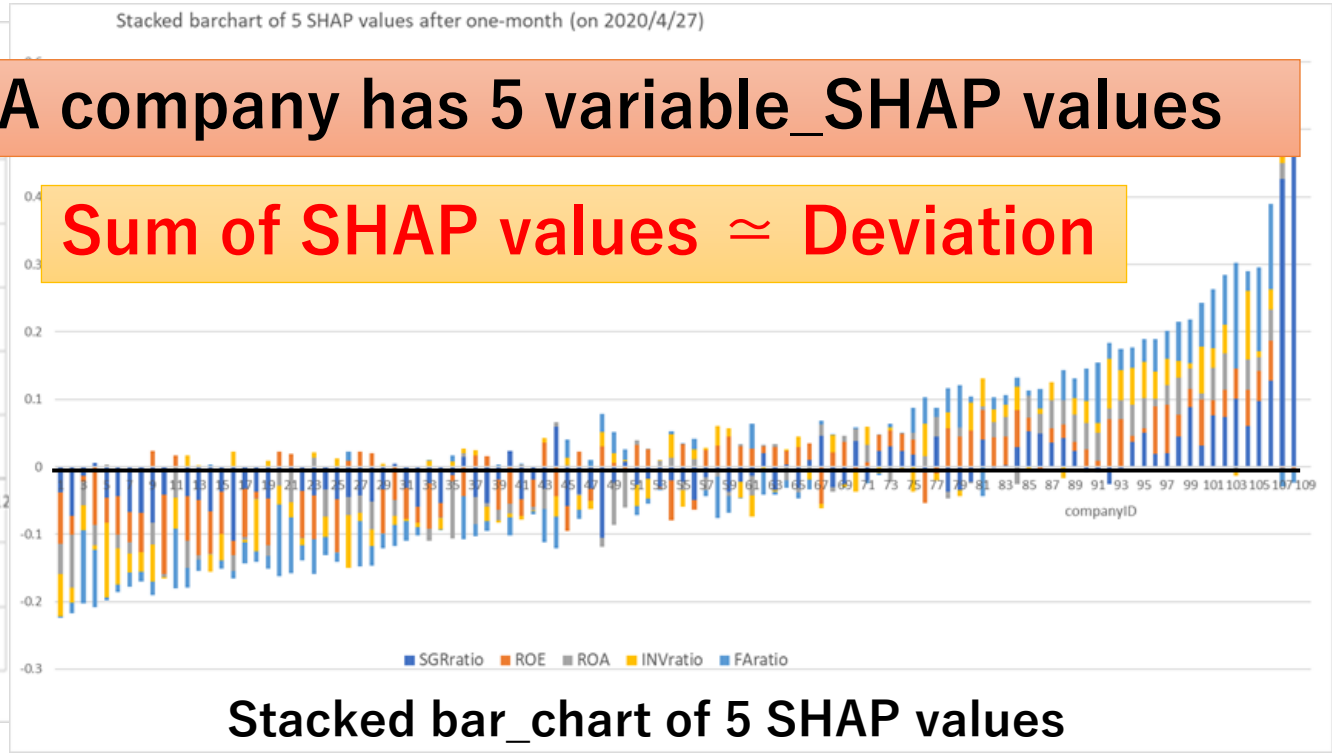
Deviation is **target - average**



108 companies

A company has 5 variable_SHAP values

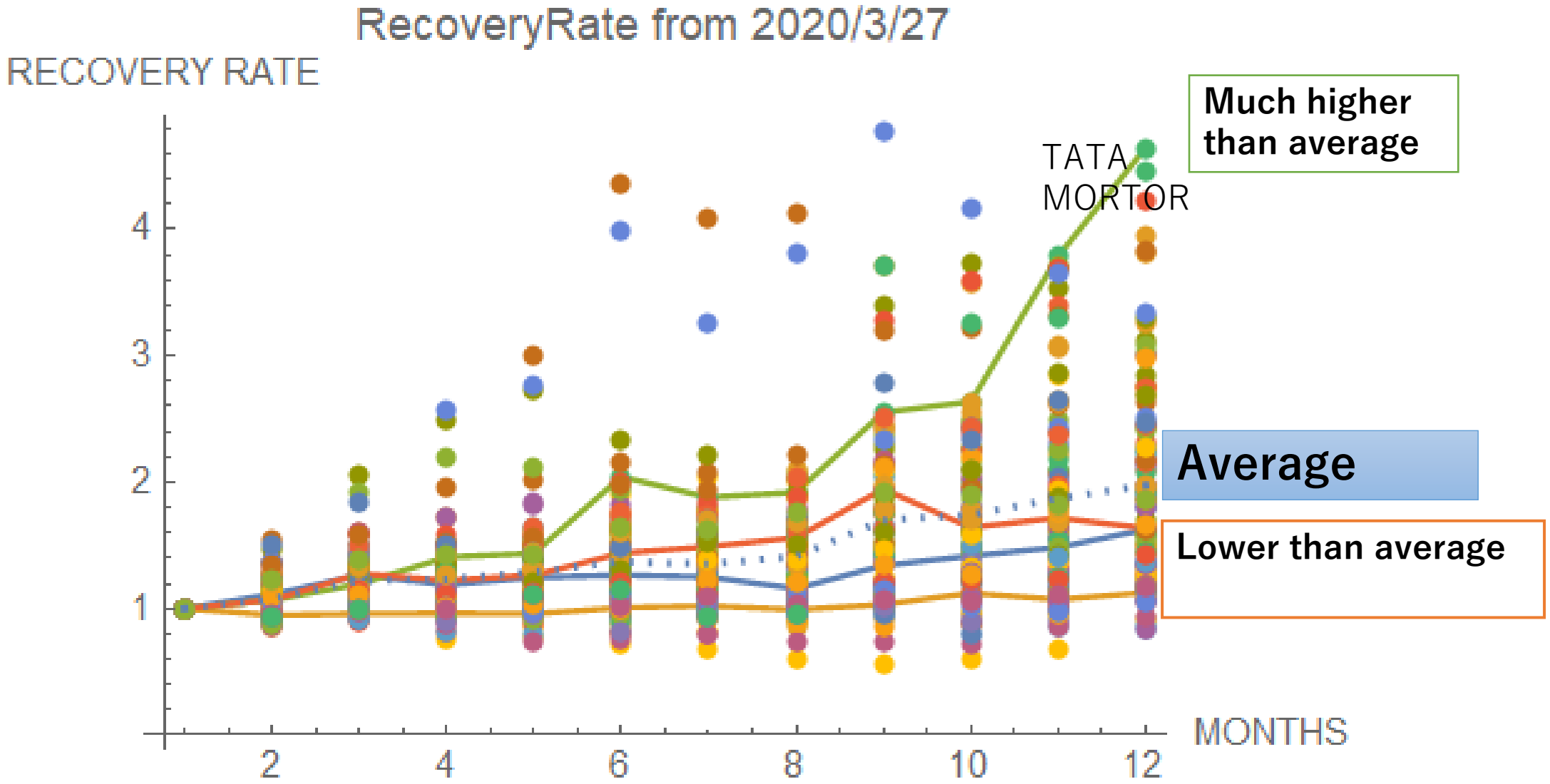
Sum of SHAP values ≃ Deviation



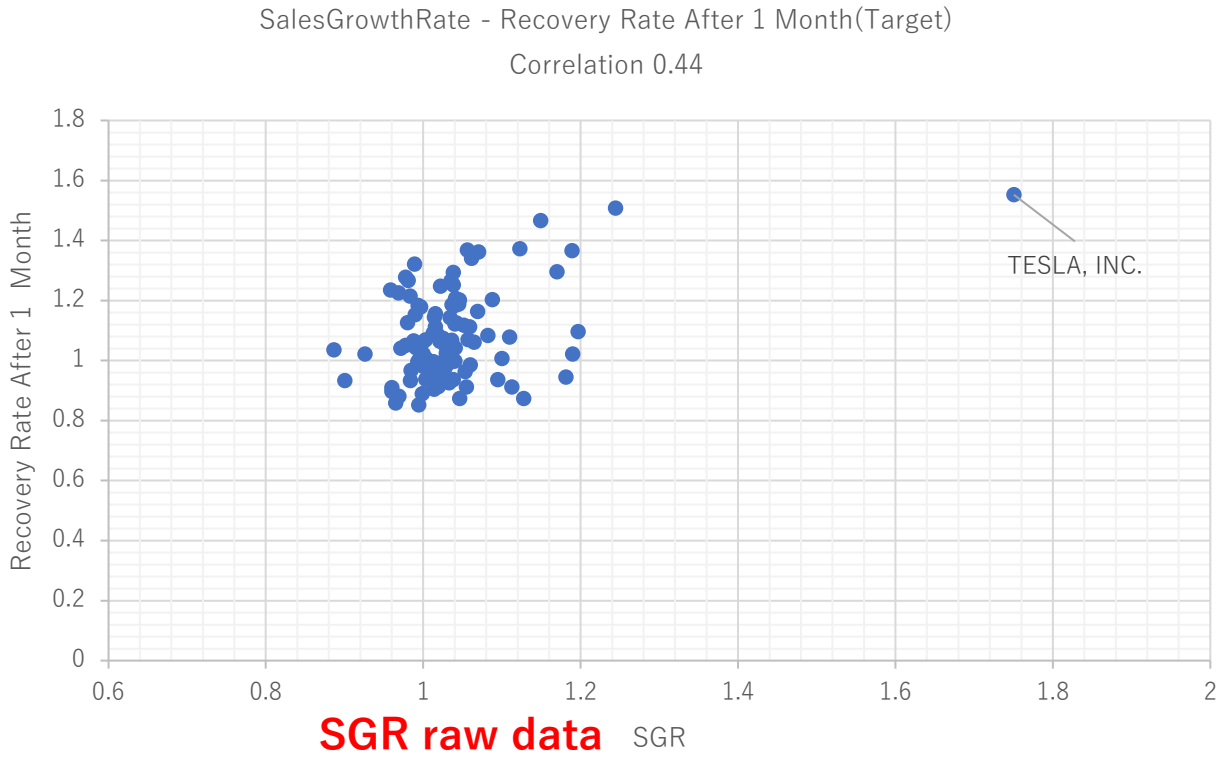
Stacked bar_chart of 5 SHAP values

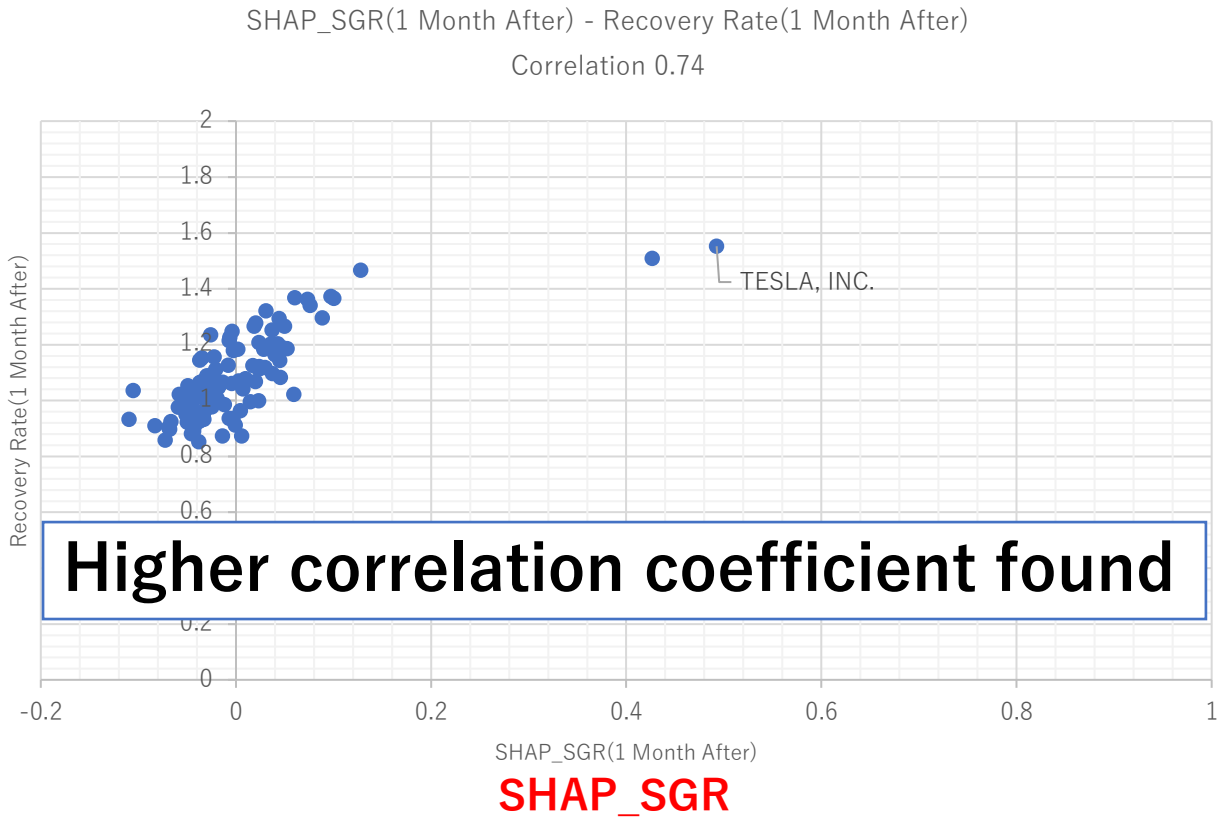
- In each company, sum of 5 SHAPs becomes target deviation

108 Companies' Recovery Rate Movement



Correlation with SHAP_SGR and target

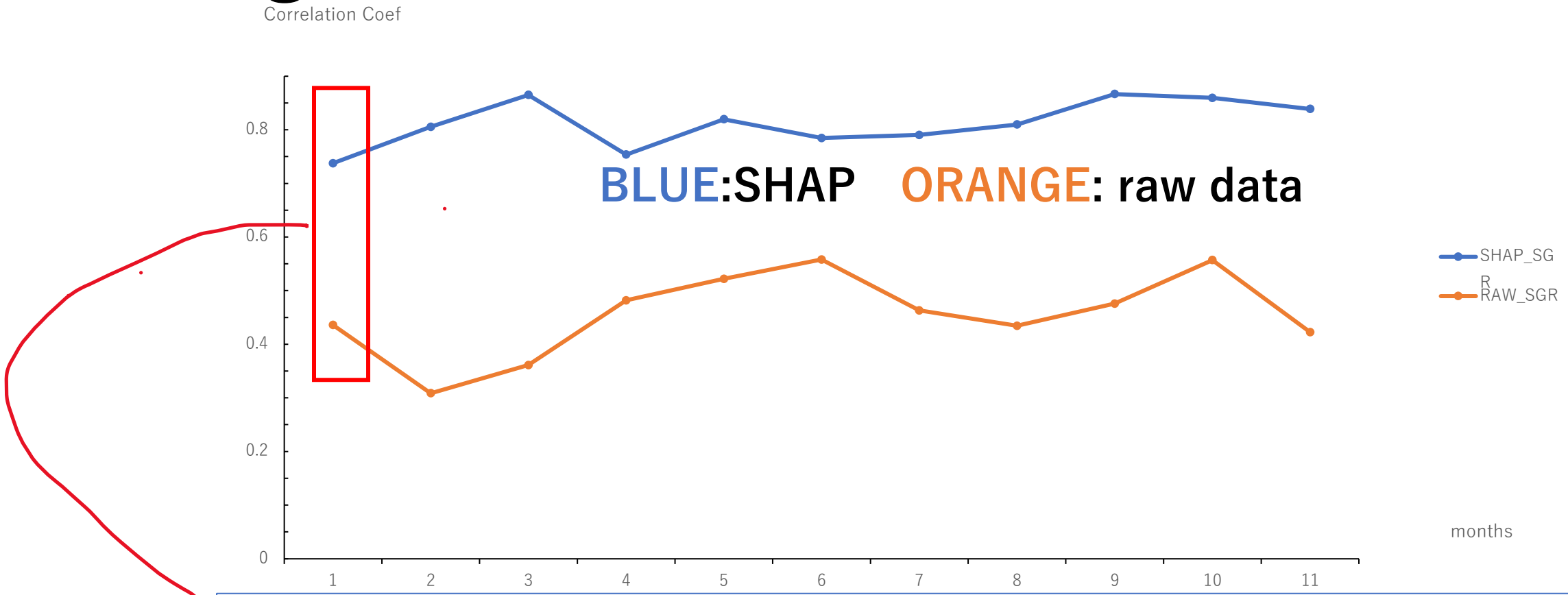




Correlation Coef. 0.44→0.74

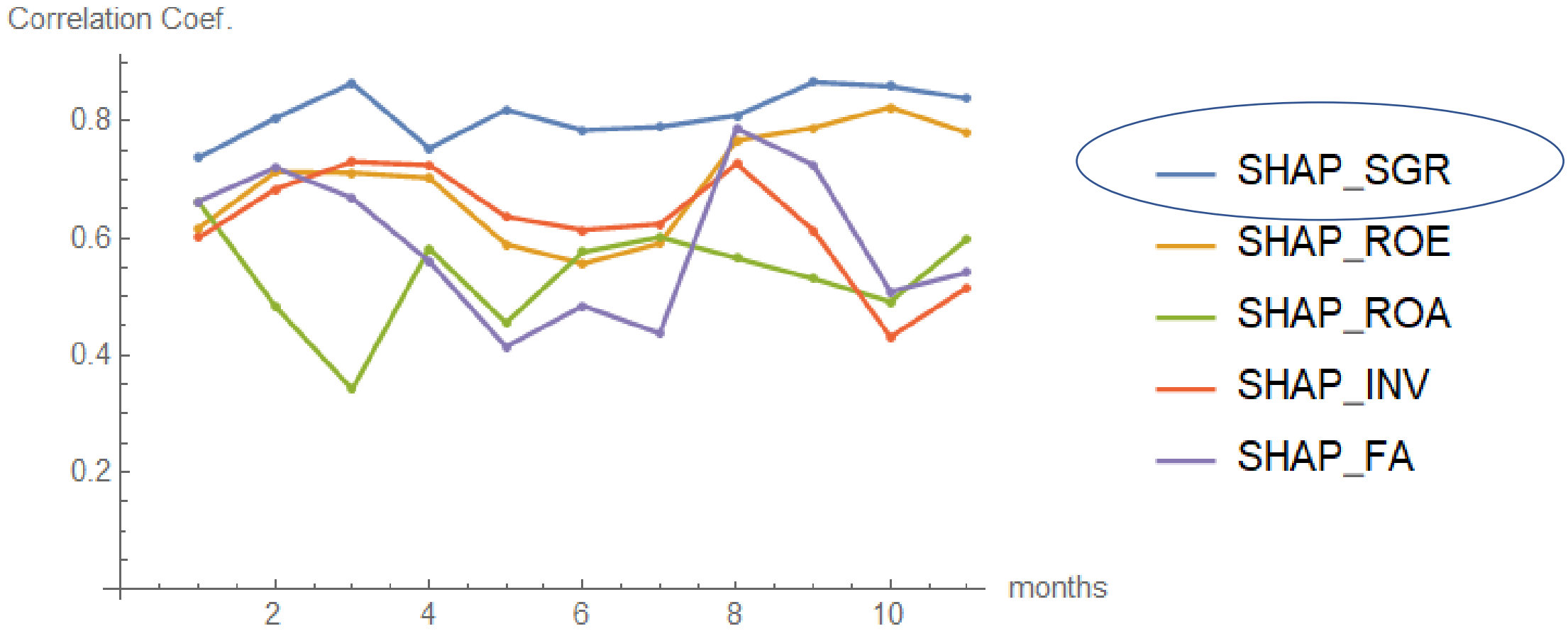
SGR contributes to the increase of Recovery Rate

Correlation with SHAP_SGR and target During 11 months



Correlation Coef. 0.44→0.74

Which is the most important factor ?
Which SHAP_variable is the highest correlation ?

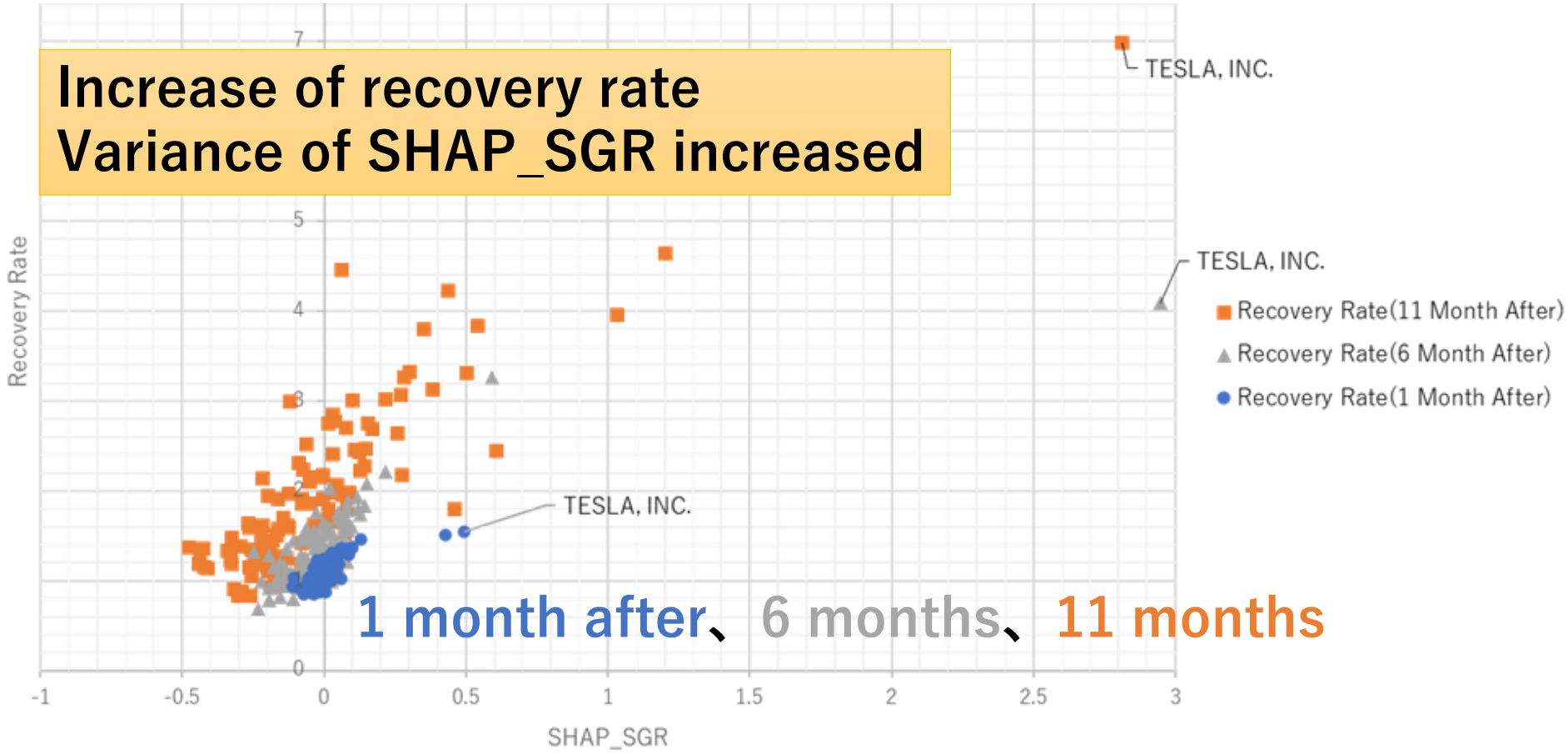


SHAP_SGR is the most important factor.

SGR is the most important factor.

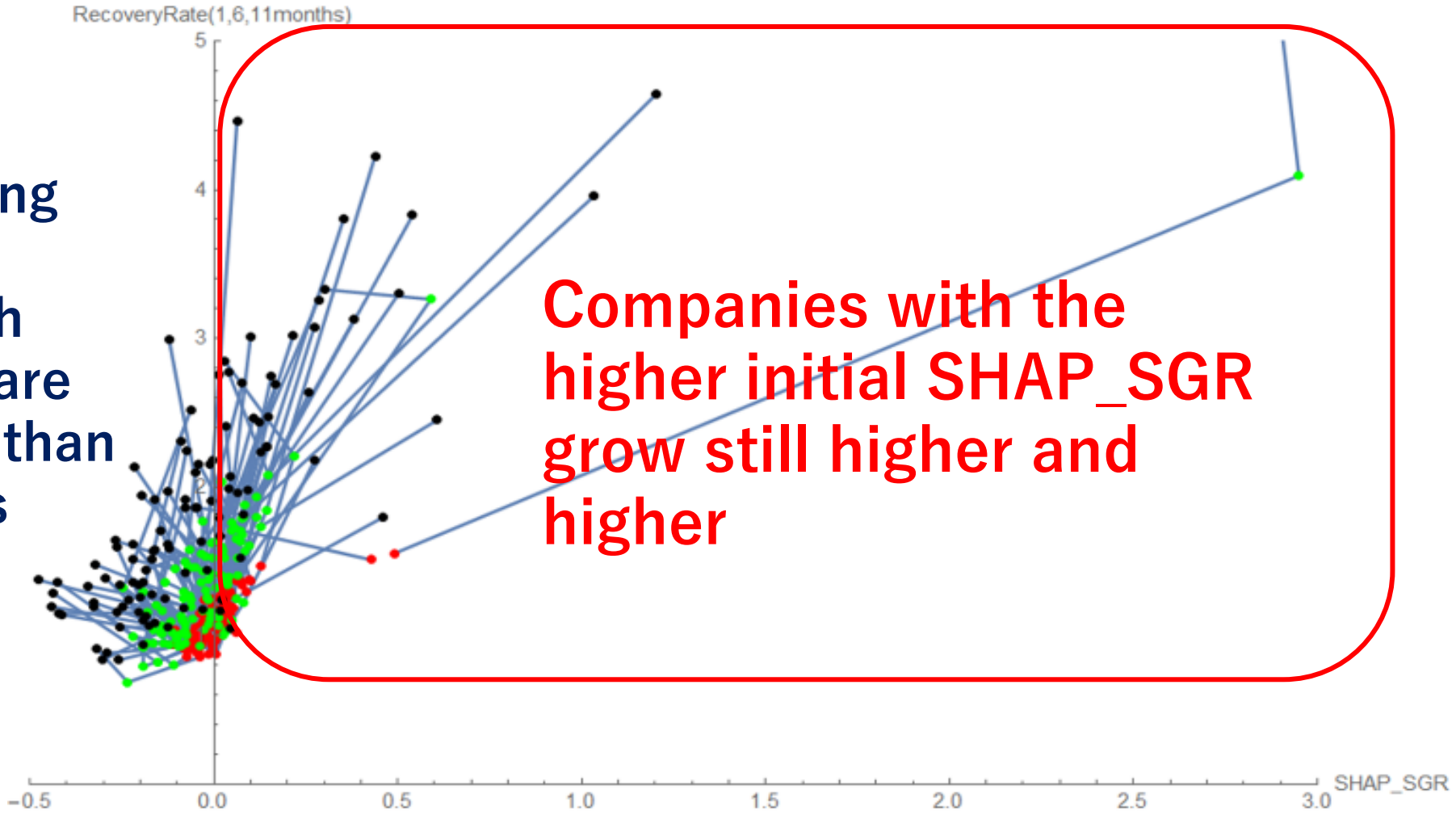
SHAP_SGR time series analysis

- How relationship was changed between SHAP_SGR and target？



SHAP_SGR time series analysis of each company

Growing but growth rates are lower than others



Companies with the higher initial SHAP_SGR grow still higher and higher

Conclusion

- Global automakers stock recovery rate at COVID-19 outbreak
- Regression analysis with SHAP values
- 11 Months Time Series SHAP Analysis

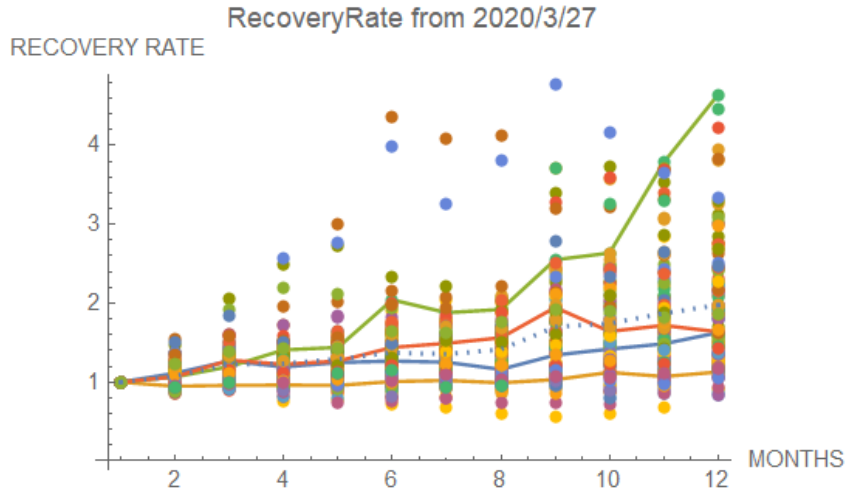
① Sales Growth Rate is important

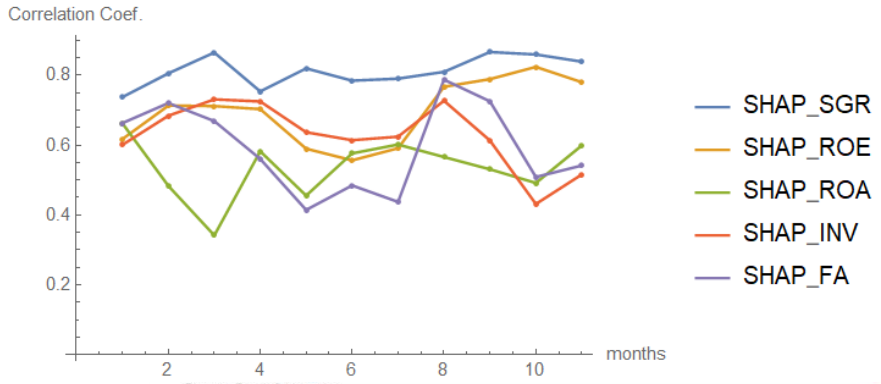
SGR__SHAP has the highest correlation

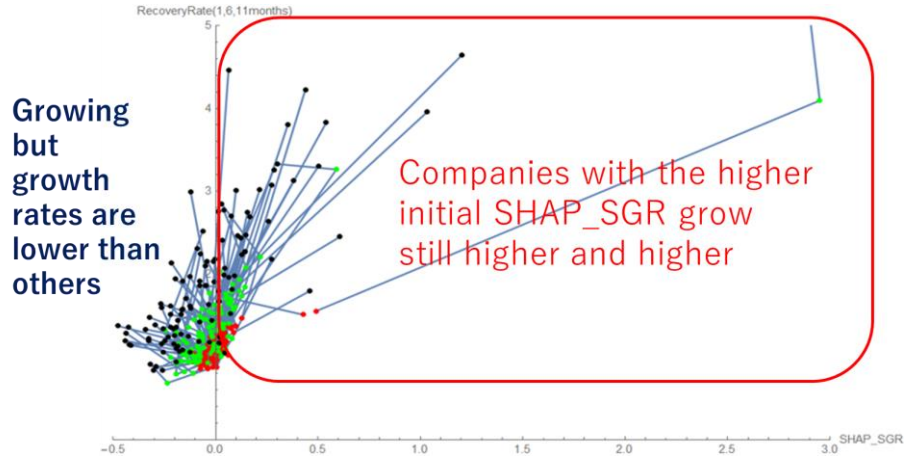
Through 11 months

② Companies with the higher initial SHAP_SGR grow still higher and higher

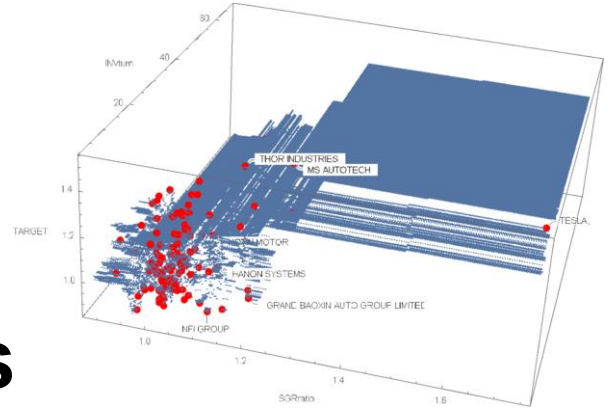
SHAP approach is applicable to many fields.







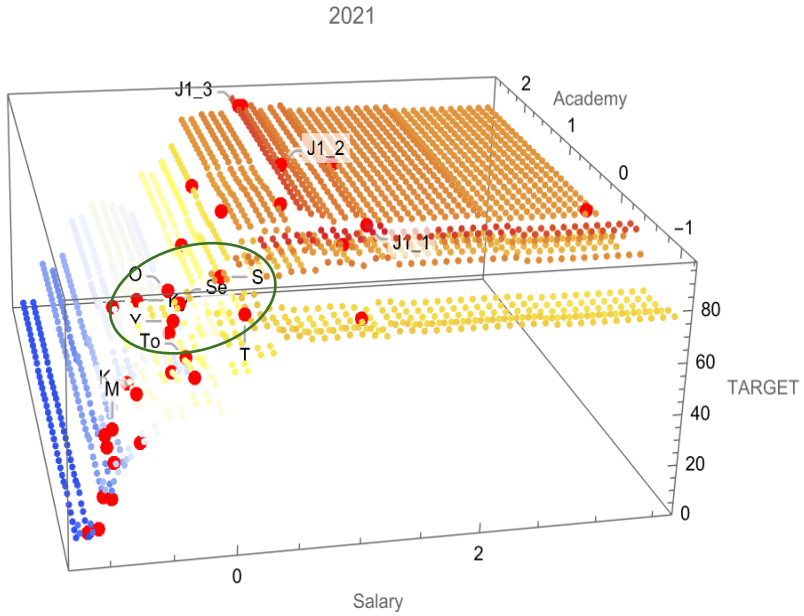
Contents

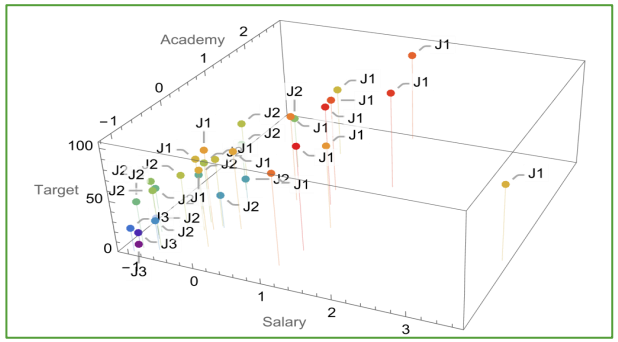


1. Graphical explanation of Shapley values
2. Cooperation game by explanatory variables
3. Theory of Shapley values
   A) Formula of Shapley values
   B) Case of bivariant
4. Case1: Time Series Analysis of SHAP Values by Automobile Manufacturers Recovery Rates
5. Case2: Football Teams Sustained Growing by Academy Training
   - Proposal of Shapley-based Measurement –
6. Conclusion

DBKDA 2023

Football Teams Sustained Growing by Academy Training - Proposal of Shapley-based Measurement -





2023/3/13 (Session 1A)

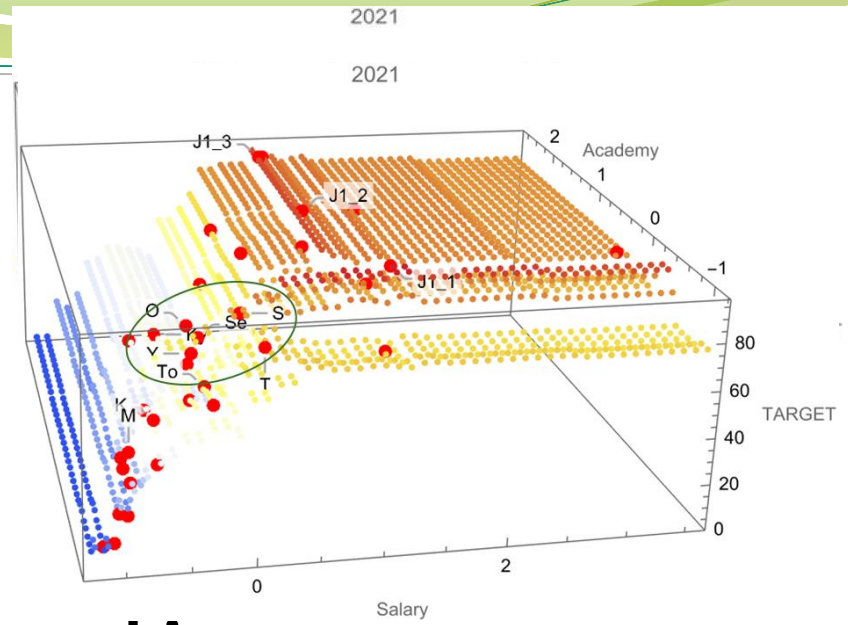
Gakushuin University

Seiji Matsuhashi

Yukari Shirota

Research Objective

- How to become STRONG football teams
- Regression with interpretation by SHAP
- Academy development is significant
  - In large-scaled teams, to sustain the high ranking.
  - In small or medium sized teams, for growth to the upper league under the limited budget.
- To measure the Academy development level, we define Matsuhashi's Measure using SHAP values.



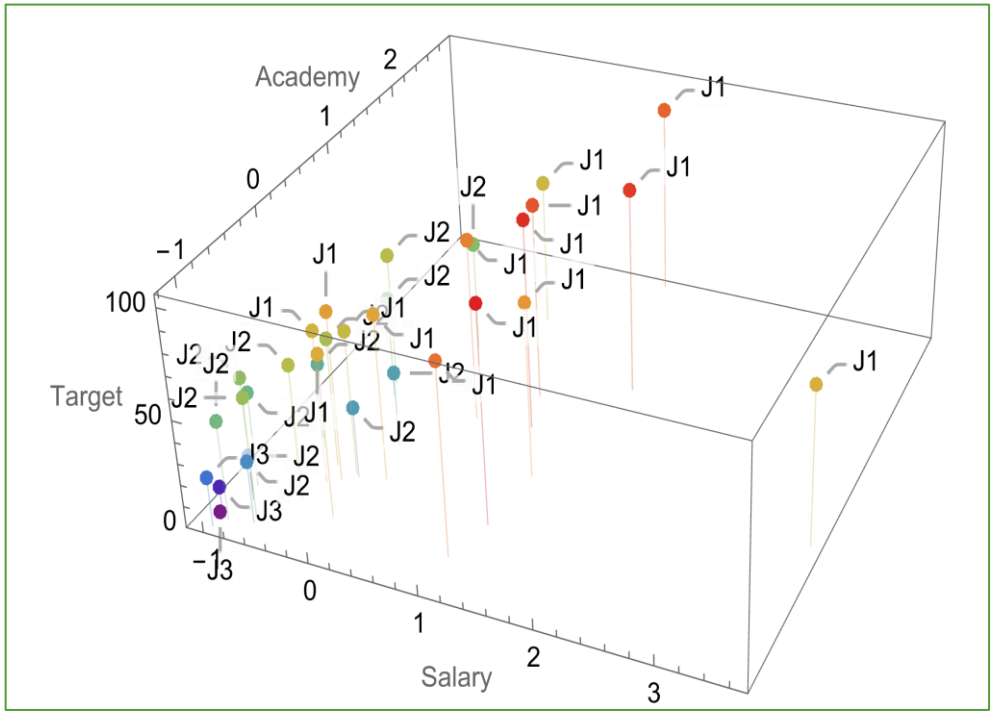
In general, SHAP can be used as KPI (Key Performance Index) definition

Regression

- **Annual ranking (0 to 100)**
  - ➢ **Time series SHAP analysis(2019 – 2021).**

Data: J1, J2, and J3

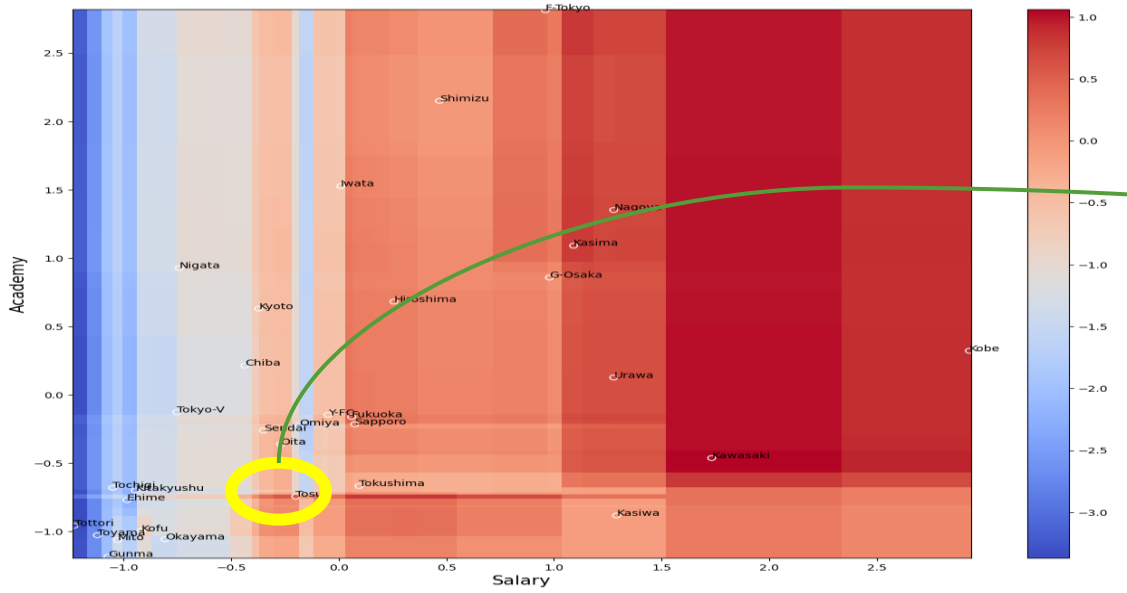
- **Explanatory variables**
  1. **Salary costs:**
     Personnel costs for the year.
  2. **Academy operating costs:**
     Total costs for 7 years.



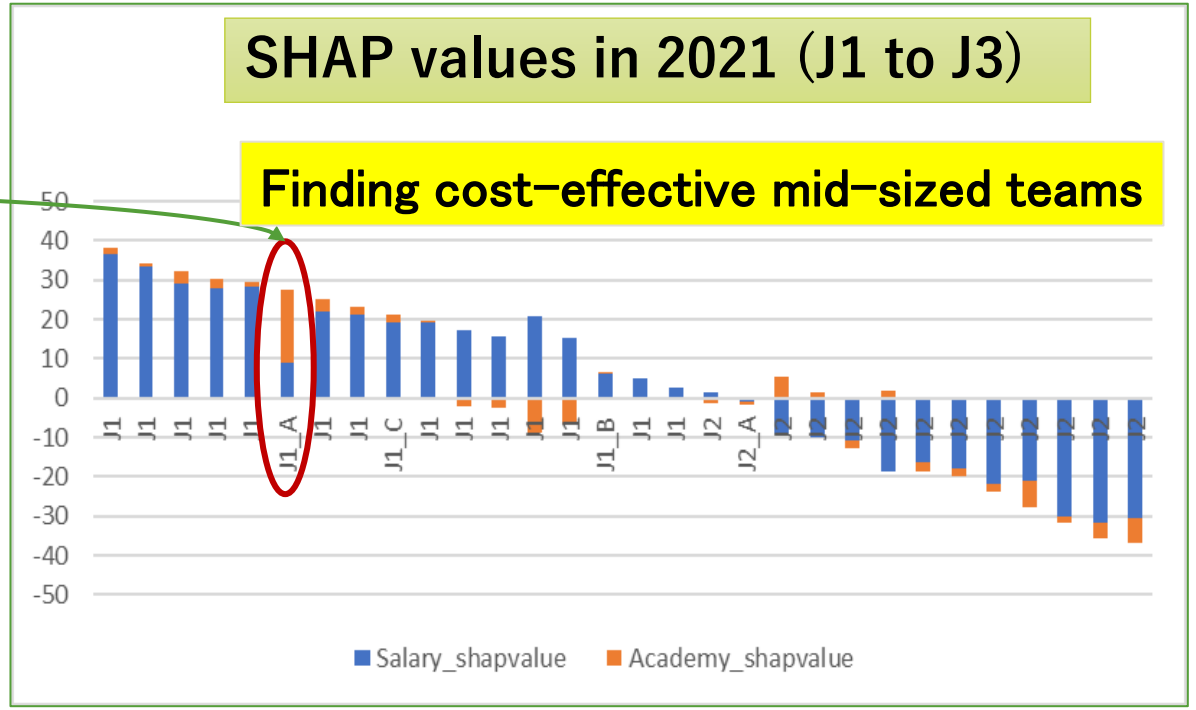
2020 input data for regression

Regression result and SHAP values in 2021

- TOSU（J1_A) has a very high Academy_SHAP value.
- TOSU : High performance under the limited budget.
- The academy graduates appearance in J1 League was also highest .



Regression model heatmap
X and y: Raw variable, not SHAP values



Measurement of Academy Development (KPI: Key Performance Index)

Matsuhashi's Measure =

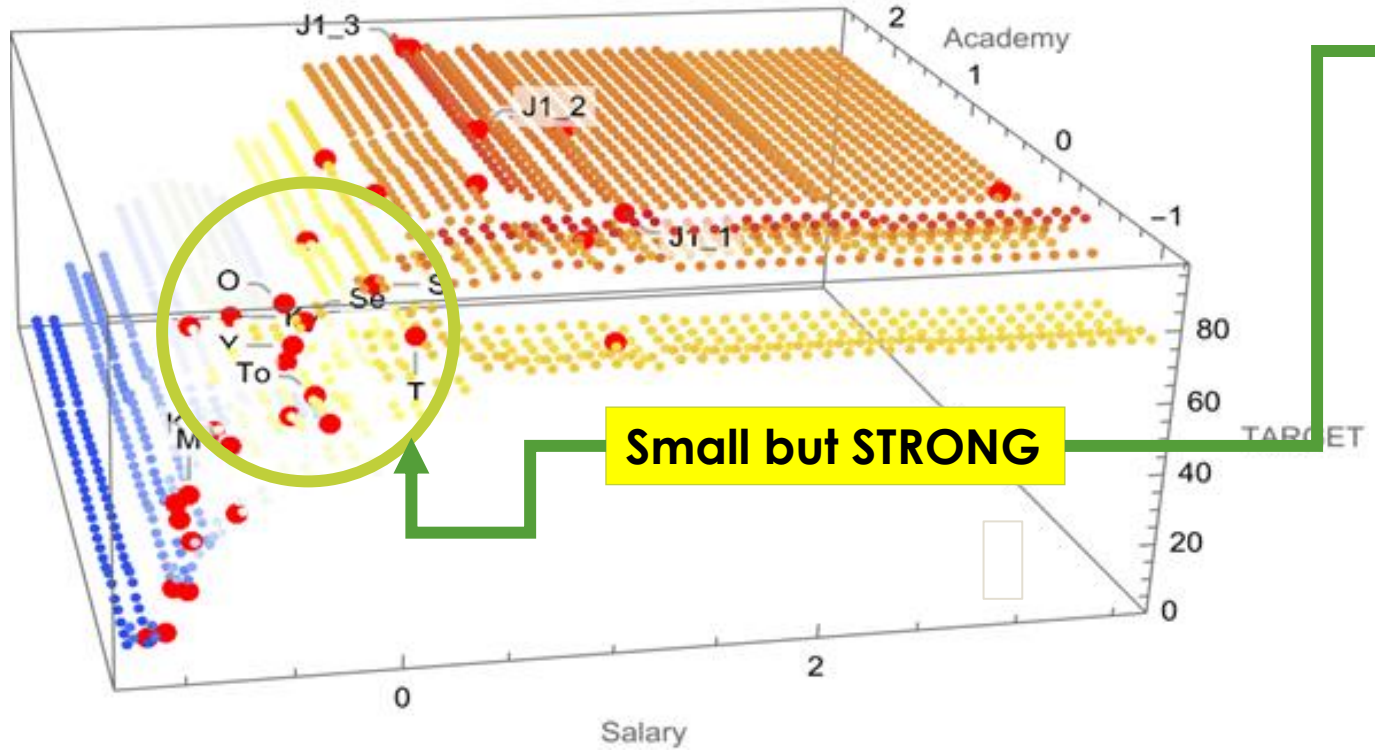
"% of Academy graduates' participant ratio" **×**

Academy_SHAP

Even if the Academy's operating costs are large, if the Academy does not generate results, the ranking score does not increase

J1-J3 (3 year-ranking)

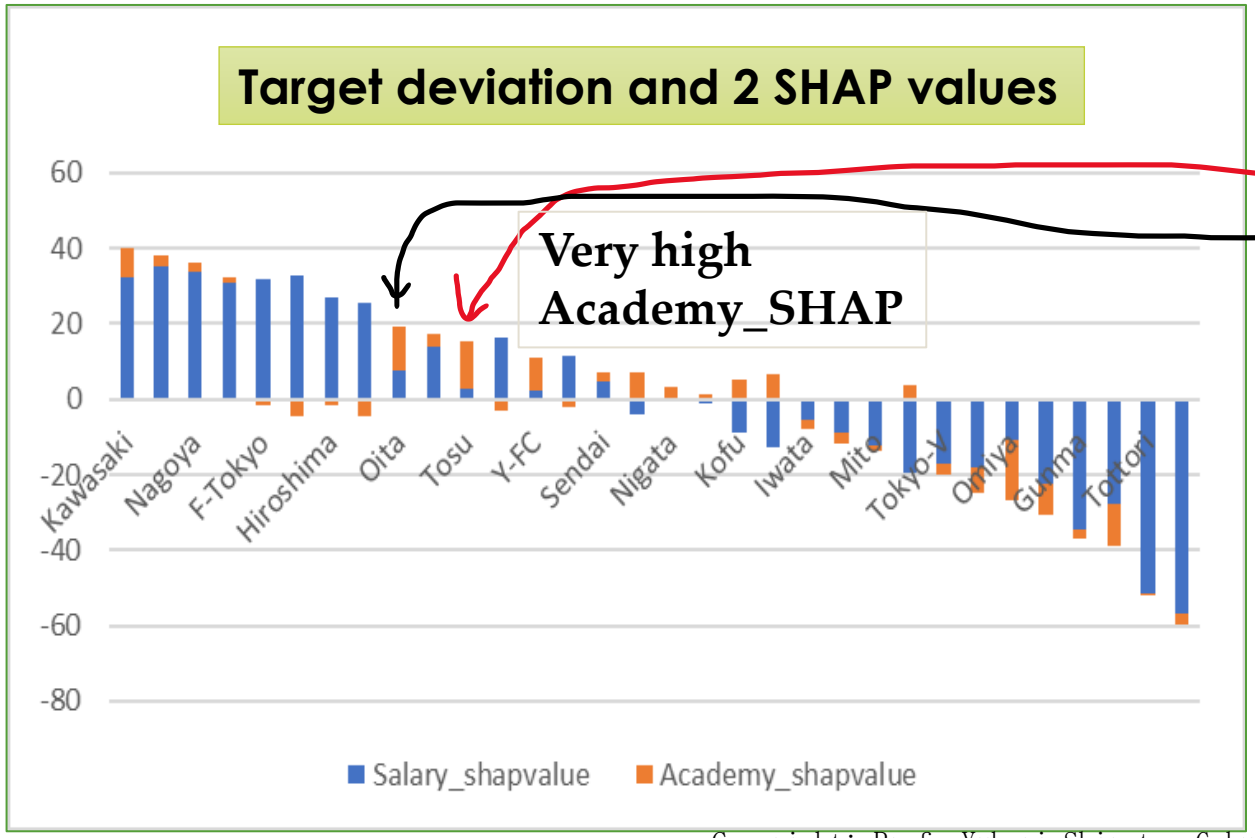
- MM can extract STRONG teams

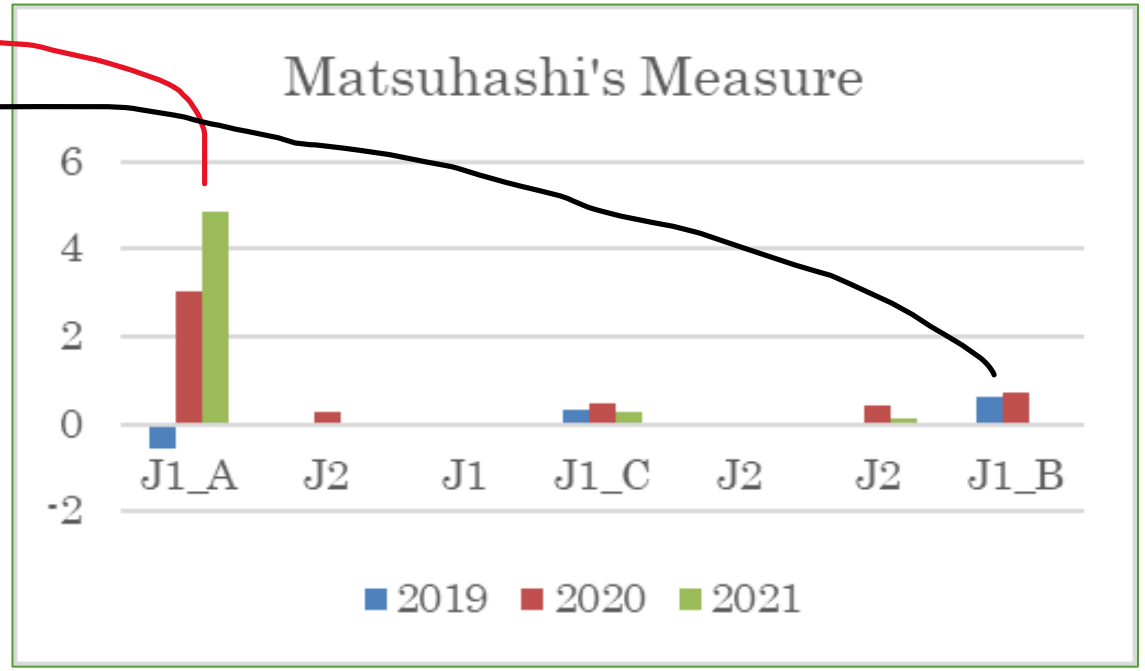


Name	Matsuhashi's Measure ('19-'21)
S	2.074
J1_1	1.540
Y	0.842
O	0.677
T	0.323
Se	0.283
J1_2	0.091
J1_3	0.088
K	0.088
To	0.018
M	0.005

Middle-sized but STRONG Teams 2020

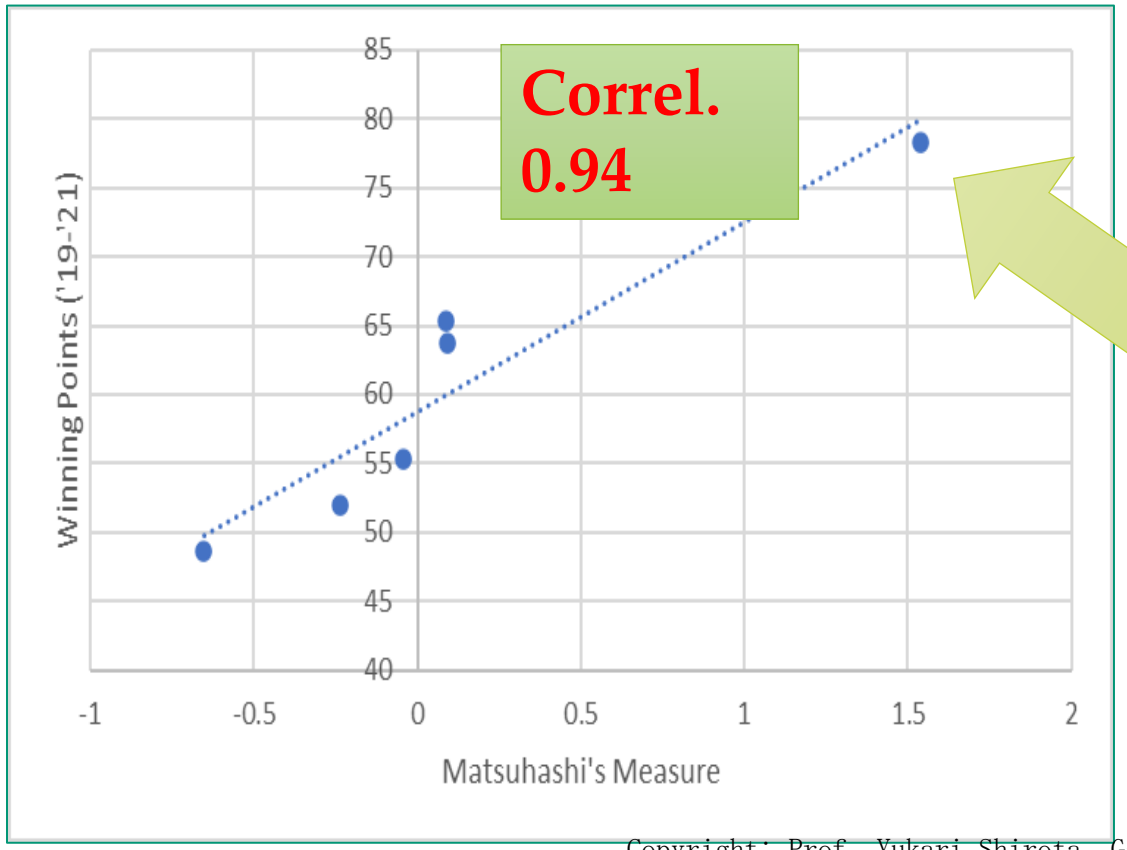
Matsuhashi's Measure Evaluation Result





Revenue TOP 6 Large Teams' MM Evaluation Sustainability of high ranking

Relationship between Matsuhashi's M and Winning Points ('19-'21)



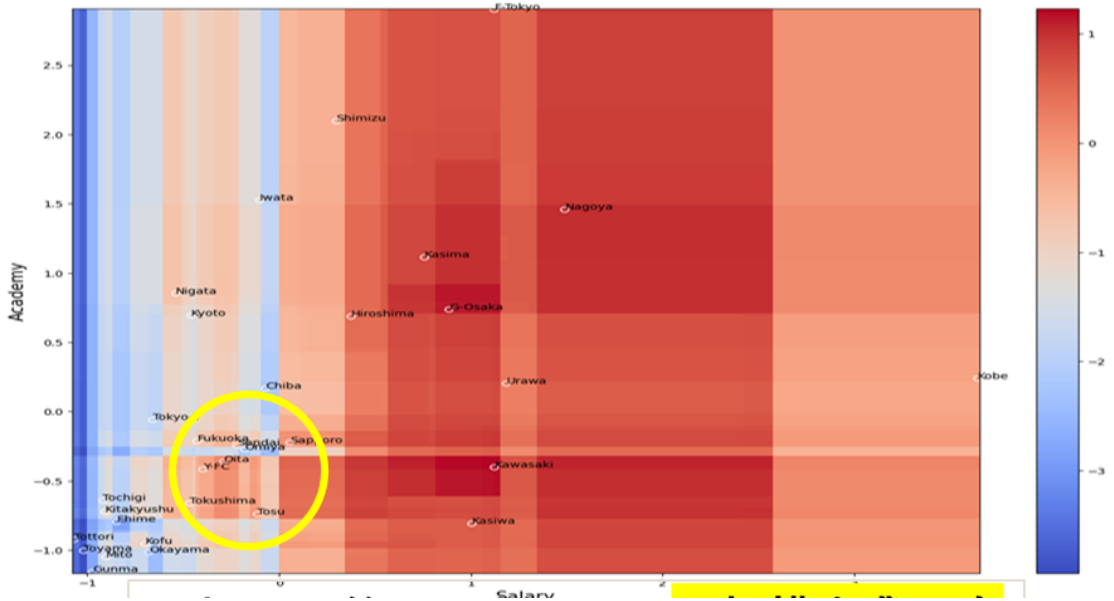
- "Revenue TOP 6 in J-League"
- High correlation with MM
  - Academy development has tight relation to the performance

Team Kawasaki
Mitoma, Tanaka,(Itakura, Kubo)
Nice performance in WC

Football Teams Sustained Growing by Academy Training

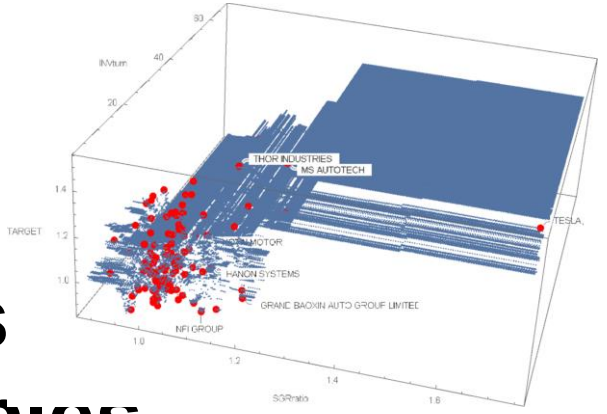
Conclusion

- **Academy development measurement: MM based on SHAP**
- **High correl. with winning**



- **In general, SHAP is effective to define KPI**
- **What are outstanding characteristics of small but high-performance companies?**
- **What is the secret of sustainable large companies?**
- **Method: SHAP is effective**

Contents



1. Graphical explanation of Shapley values
2. Cooperation game by explanatory variables
3. Theory of Shapley values
   A) Formula of Shapley values
   B) Case of bivariant
4. Case1: Time Series Analysis of SHAP Values by Automobile Manufacturers Recovery Rates
5. Case2: Football Teams Sustained Growing by Academy Training
   - Proposal of Shapley-based Measurement –
6. Conclusion

Advantage of SHAP

- In AI regressions, SHAP approach widely used
- After regression analysis,
  applicable to all application fields
- Each companies' characteristics should be evaluated
- SHAP: In the company's behavioral structure,
  each predictor's contribution
  to target can be evaluated

Stacked bar_chart is the best to present SHAP values.