Performance Analysis of Internal Corporate Ventures using Machine Learning Regression Trees and Model Trees

Mamoru Uehara

Abstract

Several studies have explored to clarify the factors that affect the performance of internal corporate ventures and the relationship between these factors and the performance. In conventional research, multiple regression analysis is often used as a method for clarifying the relationship between various factors and performance. However, as the number of factors increases, the number of multiple regression models to be examined increases, and an appropriate model must be selected from these models, in which is a time-consuming process. In addition, it is necessary to consider interactions in which factors have multiple influences, which has the disadvantage of increasing complexity. Therefore, this study attempts to overcome the abovementioned drawbacks by employing the regression tree and the model tree―extensions of the decision tree for numerical prediction―using machine learning. We then present that since the model tree uses the data classification capability of decision trees, it is possible to build a linear regression model for each classified rule and derive a model that is closer to the actual situation.

Keywords: machine learning; decision tree; regression tree; model tree; internal corporate venture; causation & effectuation; venture performance

1. Introduction.

Internal corporate ventures (ICVs) are new businesses that are established and grown within an existing company (parent company). ICVs are viewed as a means to acquire knowledge, develop new capabilities, and stimulate growth and profitability in new business areas by leveraging the resources of the parent company (Garvin, 2004). ICVs often target business domains that are new to the parent company. However, the pursuit of novelty exposes ICVs to uncertainty, which often renders it difficult for managers of both the parent company and the ICVs to accurately and completely predict the challenges they may

encounter (Covin *et al.*, 2018). Parent company managers and ICV managers must deal with this uncertainty while improving the ICV performance.

Considering these issues, it is essential to clarify the factors (venture planning autonomy, venture goal evolution, initial venture proposition clarity and learning proposition, etc.) that affect the performance of ICVs and the relationship between these factors and the performance, and many studies have been conducted on this front. Such studies often use multiple regression analysis to clarify the relationship between the various factors and the performance of ICVs. However, when the number of factors increases, hundreds of multiple regression models may be created, from which an appropriate model is selected. This takes a considerable amount of time, and in some cases, it is necessary to consider interactions in which factors have multiple influences, which has the disadvantage of increasing complexity. Therefore, it is necessary for a method to automatically select factors, even if the number of factors is large, without the need to specify or assume a model beforehand. This study attempts to overcome these shortcomings of multiple regression analysis through regression trees and model trees, which are decision trees for predicting numerical values, by using machine learning to clarify the factors that affect ICV performance.

2. Prior Research

2.1 Factors Affecting ICV Performance

Several studies have been conducted on the factors that affect the venture performance (VP) of ICVs, and the relationships between them. This section reviews some of these studies.

2.1.1 Venture Planning Autonomy and Venture Goal Evolution

Using data collected from 145 ICVs, Covin *et al.* (2019) found that when the goals of these ICVs were stable throughout the venture, the venture planning autonomy (VPA) (defined as "the extent to which the venture's management team is responsible for establishing goals, timetables, and strategy for the venture" (Johnson, 2012, p. 2)) was related to the VP. They also indicated that the venture goal evolution (VGE) had an effect at a 1% significance level; however, VGE was not significantly related to the venture value proposition evolution (VVPE). Furthermore, between the VP and the interaction term VPA×VGE was negative and significant ($p < 0.05$).

VPA implies that the decisions related to venture planning are made at the level of the ICV manager rather than the upper levels of the parent company, which places the

responsibility and authority for decision-making in the hands of those best positioned to directly observe and manage the ICV.

Covin *et al.* (2019) indicated that given that ICVs were novel ventures for the parent company, it was generally or even necessarily unreasonable to assume that the parent company's involvement in the planning of the ICV would add significant value to its operation. In addition, Covin *et al.* stated that the parent company being unfamiliar with the operation of the ICV was generally a rationale for advocating that new ventures be allowed planning autonomy. Furthermore, Covin *et al.* (2019) noted that for a parent company to enter a new domain and produce a successful ICV, the ICV manager is required to learn through experimentation, which often requires discretion and autonomy for the ICV manager. Moreover, they mentioned that as a consequence of the experimental nature of ICVs, ICV managers often pursued strategies that require a high degree of decision-making flexibility, and that autonomy allowed ICV managers to adapt and customize their ICVs to the competitive domain. Therefore, Covin *et al.* (2019) stated that VPA provided ICV managers the flexibility to implement new plans and decisions as the ICV domain became more familiar. Thus, it is of utmost importance that the VPA is such that the decision-making leadership is strategically delegated to those who have the most visibility into the matters related to the ICV business and competitive positioning, that is, the ICV managers (Covin *et al.*, 2019). Subsequently, using data from 145 firms, Covin *et al.* (2019) showed that the VPA promoted the VP.

In addition to Covin *et al.* (2019), several other theoretical and empirical studies pointed out that VPA is positively related to ICV performance. In contrast, some previous studies argued that if VPA is too high, the VP of ICV decreases, and the observations and claims observed are not uniform. For example, Johnson (2012) conducted a study of 38 ICVs and reported that VPA negatively impacted VP. These studies argued that it was necessary to strike a balance between autonomy, control, and supervision by corporate management to extract the maximum value from the new ventures.

2.1.2 Initial Value Proposition Clarity

The value proposition of an ICV defines the basis on which the ICV appeals to the market, thereby creating demand for its products and services. Covin *et al.* (2015) indicated through empirical analysis that there was a positive correlation between the VP and the initial value proposition clarity (IVPC). Kuratko *et al.* (2009) also identified a significant correlation between the VP and IVPC, noting that if the value proposition of the venture was correct from the onset, the venture profitability was generally best maintained. In addition, Covin *et al.* (2018) suggested that the clarity of the value proposition at the

beginning of the ICV establishment may have had a lasting impact on the ICV success, along with the degree to which the ICV IVPC had enhanced learning proficiency and knowledge and promoted its VP. The authors also suggested that learning had a positive impact on the success of entrepreneurial activities in the context of high environmental uncertainty. Specifically, they conducted an empirical analysis and explained that the ICV IVPC influenced the level of learning proficiency and knowledge of the ICV, which in turn influenced its VP. The results showed that the VP of the ICVs that had high IVPC were hardly affected by the level of the learning proficiency, whereas ICVs with low IVPC were able to increase their VP with high learning proficiency (achieving the same level of VP as ICVs with a high IVPC).

2.1.3　Learning Proficiency

In the field of new businesses, ICVs are required to demonstrate learning proficiency in the process of business development. However, the learning proficiency of ICVs may affect the VP differently depending on the approach and development of the various aspects of the ICVs' business plans.

Covin *et al.* (2018) employed data from 145 manufacturing ICVs in Midwestern, United States and found that the ICV learning proficiency was significantly related to the ICV VP. The VP here was based on the VP measure presented by Johnson (2012) because measuring profitability is a limitation in ICV research. In other words, the VP was captured on a qualitative and subjective scale considering four points: (a) whether the parent company's expectations had been met, (b) whether the parent company considered the ICV to be entirely successful, (c) whether the parent company believed that the ICV had achieved its planned milestones, and (d) whether the ICV performed well in terms of the achievement criteria set by the parent company. In addition, Covin *et al.* (2018) suggested that the ICV learning proficiency affected the VP when the ICV goals showed insignificant evolution during the venture development process (that is, when goal evolution was low).

2.1.4　Venture Opportunity Identification Mode

Garrett & Covin (2015) pointed out that ICVs whose domain was a field related to the parent company business could access and utilize the relevant knowledge of the parent company. Furthermore, it was indicated that the ICV VP was based on both the knowledge-based resources (stock) and the knowledge exchanged between the parent company and ICV (flow). In their study, however, Garrett & Covin consider the venture opportunity identification mode (VOIM) when the ICV's founding is based on well-considered data and information, and the parent company carefully plans (or does not plan)

the ICV's entry into the intended market. They also suggest that when the ICVs are founded in the parent company's familiar domain, well-considered business-related knowledge is incorporated into the establishment of the ICV. In other words, the discovery of business opportunities led by the parent company can be regarded as VOIM. On the other hand, Garret & Covin (2015) also suggest that ICVs established in domains not closely related to the parent company's core business are unlikely to be provided with extensive knowledge from the the parent company regarding their target market. For the former (ICVs whose domain was a field related to the parent company business), the problem is properly structured and documented (formalized processes are in place), and knowledge can be leveraged in known or predictable ways. The ICV can also take over the organizational knowledge accumulated in the parent company. This knowledge is an inherent asset of the firm that cannot be easily imitated; furthermore, it is not tradable.

2.1.5 Causation and Effectuation

To handle the uncertainty associated with establishing a new venture, entrepreneurs can choose from a variety of strategies. In an attempt to address this central research question in entrepreneurship, Sarasvathy (2001) proposed effectuation as the dominant decision model for entrepreneurial decision-making, particularly when there is no existing market.

Sarasvathy (2008) explained effectuation as follows:

> "Effectuation is the inverse of causation. Causal models begin with an effect to be created. They seek either to select between means to achieve those effects or to create new means to achieve preselected ends. Effectual models, in contrast, begin with given means and seek to create new ends using non-predictive strategies. (Sarasvathy, 2008, p. 16).

Sarasvathy (2008) then explained that causation-based strategies were effective when the future was predictable and the goal was clear, whereas effectuation-based strategies were effective when the future was unpredictable and the goal was unclear.

Sarasvathy (2008) also pointed out that empirically, entrepreneurs utilized both causation and effectuation approaches in various combinations, and the preferred mode usually depended on the entrepreneur's degree of expertise and the stage of the life cycle of the firm.

Although several previous studies have suggested that VP is related to both causal business planning and effectual action orientation, the potential synergy between the two logics has not been studied. Thus, Smolka *et al.* (2016) investigated and tested the mutual

relationship between causation and effectuation for VP. They empirically verified that entrepreneurs frequently used a combination of causation and effectuation, that there is an interrelationship between effectual and causal decision-making, and verified the interactions between the two and their VPs. As a result, Smolka *et al.* (2016) confirmed that the entrepreneurs' use of causal and effective reasoning in combination had a positive effect on the VP.

2.2 Measurement Scale for Causation and Effectuation in ICV

Chandler *et al.* (2011) developed and validated measures of causation and effectuation―the decision-making processes of entrepreneurs―and tested it using a sample. The developed measurement of causation consisted of well-defined, coherent, and unidimensional elements. The measure of effectuation, however, was shown to be a formative, multidimensional construct with three related sub-dimensions (experimentation, affordable loss, and flexibility) and one dimension (composed of pre-commitments) in common with the measurement of causation. According to Sarasvathy (2001), causation is negatively related to uncertainty, whereas the factor experimentation, a sub-dimension of effectuation, is positively related to uncertainty. The scale of Chandler *et al.* (2011) measured causation and effectuation through a 20–item questionnaire rated on a five-point Likert scale. Thus, the significant contribution of the Chandler *et al.* (2011) study was to validate a scale for the measurement of causation and effectuation.

2.3 Regression Tree and Model Tree

Generally, regression analysis is the first preference for numerical prediction tasks; however in some cases, decision trees may be a more beneficial alternative to regression models. For example, decision trees may be more suitable for tasks with a large number of factors (independent variables) or where a number of complex relationships between factors (independent variables) and dependent variables are allowed (Lantz, 2019).

Decision trees for predicting numbers consist of two categories, one of which is the regression tree, which, contrary to the name, does not use linear regression techniques. Instead, regression trees generate predictions based on the average value of the instances that reach the leaf node (Lantz, 2019).

The second category for numerical prediction is the model tree. Although model trees grow similar to regression trees, a multiple regression model is built for each leaf node from the instances that have reached that node (Lantz, 2019).

The construction of a decision tree for numerical prediction is similar to that of a decision tree for classification: starting from the root node and dividing the data using a divide-and-conquer strategy. The divide-and-conquer strategy uses the features that provide the highest uniformity for the result. One of the most commonly used criteria for division is the standard deviation reduction (SDR), which is defined by the following equation (Lantz, 2019, Quinlan, 1992 and 1993).

$$\mathrm{SDR} = sd(T) - \sum i \frac{|T_i|}{|T|} \times sd(T_i)$$

where the $sd(T)$ function represents the standard deviation of the values in the set T, and T_1, T_2, ...T_n represent the set of values resulting from the feature-based division. The $|T|$ term represents the number of observed values in set T. Basically, we measure how much the standard deviation decreases by comparing the standard deviation before the split to the weighted standard deviation after the split (Lantz, 2019, Quinlan, 1992 and 1993).

Another way to study the performance of a model is to find out how far, on average, the predictions are from the correct answer. This index is called the mean absolute error (MAE), which can be expressed as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

where n is the number of predictions, and e_i is the error of prediction i (Lantz, 2019). As the name MAE implies, this formula finds the mean of the absolute values of the errors.

3. Factors Affecting the VP of the ICV and their Measurements

This section describes the method of measuring the VP and the factors affecting the VP of the ICVs examined in this study. First, we describe how the VP of ICVs is measured before detailing the factors that affect the VP and their measurement methods. As described in Section 2, various factors affect the VP; however, this study utilizes the factors (VPA, VGE, IVPC, VOIM, VGE, learning proficiency, causation, effectuation, and pre-commitments) described prior studies in Section 2.

3.1 Measurement of the ICV VP

Covin *et al.* (2018) noted that using financial measures to measure the VP could be problematic because ICVs were new businesses starting with zero sales. Therefore, they

explained that subjective measures were the most reasonable approach to measure the VP in ICV research.

Johnson (2012) also pointed out that the profitability criteria were not appropriate for assessing the performance of early-stage ventures because they were not profitable and measured the VP using a subjective scale based on four statements rated on a seven-point Likert scale.

This study utilizes the VP scale employed by Johnson (2012).

3.2 Measurement of the VPA and the VGE

As mentioned in 2.1.1, Covin *et al.* (2019) studied the impact of the VPA and the VGE on the VP. This study employs the same scale proposed in Covin *et al.* (2019) to measure the VPA and the VGE.

3.3 Measurement of the IVPC

As mentioned in 2.1.2, Covin *et al.* (2015) studied the impact of the IVPC on the VP. This study used the same scale as Covin *et al.* (2015) to measure the IVPC.

3.4 Measurement of the VOIM

Covin *et al.* (2018) and Garrett & Covi (2015) argued that ICVs with domains in areas related to the parent company's operations were more likely to be successful because they had access to the parent company's knowledge stock, could make better use of the parent company's resources, and often proceeded as planned. Garrett & Covin (2015) stated that the VP of an ICV was based on both knowledge-based resources (stock) and knowledge exchanged between the parent company and the ICV (flow), and that the parent company carefully planned the ICV's entry into the intended market degree as the VOIM. They subsequently suggested that when an ICV was established in a domain familiar to the parent company, well-considered business-related knowledge was incorporated into the establishment of the ICV.

This study uses the same scale proposed by Garrett & Covin (2015) to measure the VOIM.

3.5 Measurement of the VGE

Covin *et al.* (2018) identified that the ICV learning proficiency significantly impacted the ICV VP when the ICV goals showed insignificant evolution during the venture development process (when goal evolution is low). They measured the VGE using three statements rated on a seven-point Likert scale.

This study uses the scale proposed by Covin *et al.* (2018), which assesses the extent to which venture goals change throughout an ICV's founding, as a measure of the VGE.

3.6 Measurement of Learning Proficiency

Covin *et al.* (2018) studied the relationship between the VP and the knowledge increase/learning proficiency in the founding process of the ICVs. In their study, they developed a scale for the learning proficiency. They computed the ICV learning proficiency index by multiplying scores on scale items reflecting "the adequacy of venture knowledge" in specific, carefully choosing venture management-related areas by scores on scale items reflecting "the extensiveness of knowledge acquisition" in those same areas.

Knowledge adequacy measures the extent to which knowledge is possessed on a seven-point Likert scale. Knowledge acquisition extensiveness measures the extent to which knowledge increases during the founding process of the ICV on a seven-point Likert scale.

In this study, the scale on the learning proficiency is used, which is the learning proficiency scale developed by Covin *et al.* (2018).

3.7 Measurement of Causation, Effectuation, and Pre-Commitments

Chandler *et al.* (2011) developed and validated measurement scales for causation and effectuation (the decision-making process of entrepreneurs) and tested them using a sample. As stated by Sarasvathy (2008), causation is negatively associated with uncertainty, and the factors of experimentation (experiments, tests, and field practice), which comprise a sub-dimension of effectuation, are positively associated with uncertainty.

The scale of causation that Chandler *et al.* (2011) developed was extracted as a unidimensional construct. Nevertheless, the scale of effectuation was a formative, multidimensional construct that had three related sub-dimensions (experimentation, affordable loss, and flexibility) and one dimension that was common with the measure of causation (pre-commitments). The scale used in Chandler *et al.* (2011) measures causation and effectuation and pre-commitments through a 20–item questionnaire rated on a five-point

Table 1:　Factors (independent variables) used in this study

venture planning autonomy (VPA)
venture goal evolution (VGE)
initial value proposition clarity (IVPC)
learning proficiency
venture opportunity identification mode (VOIM)
Causation
Effectuation
Pre-commitments

Likert scale.

This study utilizes the scales proposed by Chandler *et al.* (2011) to measure the causation, effectuation, and pre-commitments.

4. Analysis Method

To clarify the factors and relationships on VP of ICVs, we first performed multiple regression analysis, which is used in many studies.　Then, an appropriate model was selected by using all the factors (independent variables) and selecting variables through the stepwise method.

Next, to automatically select the factors, machine learning was employed to identify the factors and their relationship to the ICV VP through regression trees and model trees.

The factors (independent variables) used in this study are shown in Table 1. Note that the dependent variable VP and all factors (independent variables) are standardized with a mean of zero and a standard deviation of one.

5. Empirical Analysis using Application Examples

5.1　Survey Targets

(1) A web-based questionnaire survey was administered to 292 Japanese individuals with experience in ICVs.

(2) Analysis was conducted using data from 90 respondents (valid response rate: 30.8%). The 90 responses were divided into 80% training data ($n = 72$) and 20% test data ($n = 18$) for analysis using machine learning.

Table 2: Industries

Wholesale	14	Manufacturing	12	Other Services	9
Real Estate	8	Retail	8	Information and Communication	8
Construction	5	Other	26		

Table 3: Location

Tokyo	31	Osaka	12	Kanagawa	7	Hyogo	6
Hokkaido	4	Kyoto	4	Chiba	4	Other	22

5.2 Attributes

(1) Average ICV Age: 11.1 years

(2) Average Sales Amount: 471 million yen

(3) Average Capitalization: 34 million yen

(4) Average Number of Employees: 58

(5) Industries: See Table 2.

(6) Location: See Table 3.

5.3 Analysis Results

5.3.1 Multiple Regression Analysis

Multiple regression analysis was performed using the 72 training data points using the SPSS version 27 software. The correlation coefficients between the independent variables and the dependent variable are shown in Table 4. The parameters estimated using the

Table 4: Correlation Coefficients of Variables

	Correlation Coefficient								
	VP	VPA	VGE	IVPC	learning proficiency.	VOIM	causation	effectuation	Pre-commitments
VP	1.000								
VPA	−0.032	1.000							
VGE	0.430	−0.112	1.000						
IVPC	0.550	0.031	0.494	1.000					
Learning proficiency.	0.558	0.220	0.409	0.453	1.000				
VOIM	0.445	−0.242	0.425	0.239	0.155	1.000			
Causation	0.324	−0.182	0.316	0.384	0.316	0.201	1.000		
Effectuation	0.226	−0.160	0.159	0.268	0.226	0.003	0.782	1.000	
Pre-commitments	0.393	−0.053	0.309	0.396	0.367	0.191	0.759	0.737	1.000

Table 5:　Estimation Results of Parameters

	β	t-value	Significant Probability	VIF
VPA	−0.056 n.s.	−0.579	0.565	1.265
VGE	−0.020 n.s.	−0.176	0.861	1.656
IVPC	0.296**	2.734	0.008	1.571
Learning proficiency	0.372**	3.475	0.001	1.530
VOIM	0.308**	3.021	0.004	1.386
Causation	−0.109 n.s.	−0.672	0.504	3.504
Effectuation	0.041 n.s.	0.258	0.797	3.338
Pre-commitments	0.136 n.s.	0.907	0.368	3.011
Coefficient of Determination adjusted for the degree of freedom	0.469**			

**: p ＜0.01, n.s.: non-significant

Table 6:　Parameter estimation results for the models selected by the stepwise method.

		β	t-value	significant probability	VIF	Coefficient of Determination adjusted for degree of freedom
model1	learning proficiency.	0.558**	5.630	0.000	1.000	0.302**
model2	learning proficiency.	0.501**	5.517	0.000	1.024	0.428**
	VOIM	0.368**	4.049	0.000	1.024	
model3	learning proficiency.	0.370**	3.908	0.000	1.262	0.495**
	VOIM	0.316**	3.618	0.001	1.064	
	IVPC	0.307**	3.184	0.002	1.307	

**: p ＜0.01

multiple regression analysis and variance inflation factor（VIF） are shown in Table 5.

The variable selection was carried out using the stepwise method, three models were selected, and the parameters were estimated as shown in Table 6.

5.3.2　Regression Tree

The “rpart” package of R was used to build the regression tree model.　R is a cross-platform, no-cost statistical programming environment.　The rpart is an abbreviation for recursive partitioning.　To visualize the regression tree, we use the rpart.plot package.　The rpart.plot package is an easy-to-use package that outputs a high-quality decision tree.　The decision tree output using the rpart.plot package displays leaf nodes at the bottom.　The number displayed in each leaf node indicates the predicted value when the particular node is reached.

The regression tree created using the training data（$n=72$） is shown in Figure 1.

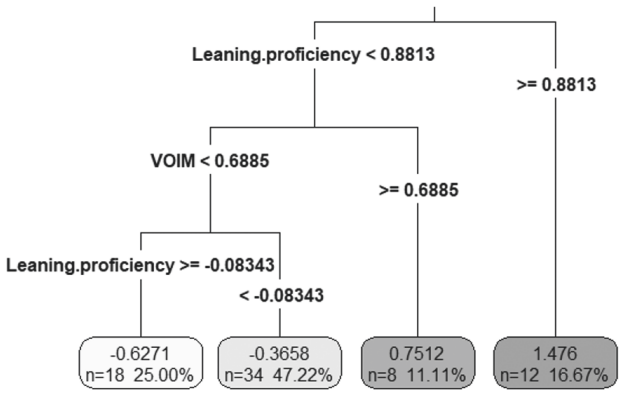

Fig. 1 Results of the regression tree analysis (n=72)

(1) For a learning proficiency ≥0.8813, the VP is 1.478.

(2) For a learning proficiency <0.8813 and a VOIM ≥0.6885, the VP is 0.7512.

(3) For a learning proficiency <0.8813, the VOIM <0.6885, and for a learning proficiency < −0.08343, the VP is −0.3658.

(4) For learning proficiency <0.8813, the VOIM <0.6885, and for a learning proficiency ≥ −0.08343, the VP is −0.6271.

The predict function in R was used to generate predictions on the test data ($n=18$). Finding the correlation between the predictions and the correct answers is a simple way of measuring the performance of the model. The "cor" function in R shows how well the predicted values correspond to the correct answers.

The correlation between the predictions and the correct answers of the test data ($n=18$) was 0.260, which is unsatisfactory. This correlation shows only the strength of the relationship between the predicted value and the correct answer value, not how far the predicted value is from the correct answer value. Therefore, the MAE as described in Section 2.3, was used to examine how far the predicted value was from the correct answer.

The MAE of the test data ($n=18$) was 0.767. This MAE (mean difference between the predicted value and the correct answer) value indicates that the model performed poorly because the dependent variable VP after standardization has a minimum value of −1.896 and a maximum value of 1.634. The MAE, which is 0.767 for the test data of the regression tree model, does not differ significantly from that of the classifier, which only predicts the mean of the training data (MAE=0.733), indicating that the model can be significantly improved.

5.3.3 Model Tree

A model tree, which extends the regression tree by replacing the leaf mode with a

regression model, was used to improve the performance of the regression tree. Model trees often provide a better performance than regression trees, which use only one independent variable in leaf mode.

The most advanced model tree at present is the "Cubist" algorithm (Quinlan, 1992, 1993, Lantz, 2019). The Cubist algorithm constructs a decision tree, creates decision rules based on the branches of the decision tree, and builds a regression model in each leaf node. Pruning and busting, which are concepts widely applicable in machine learning algorithms, are used to improve the quality of the predictions and the smoothness of the range of predictions. In R, the Cubist algorithm is provided as a Cubist function in the Cubist package.

Two rules were estimated, as shown in Fig. 2 below, in the model created by applying the R Cubist function to the training data ($n = 72$). The difference between the output of the model tree and that of the regression tree is that the model tree's nodes represent linear models rather than numerical predictions. In each rule of the model tree, the "then" followed by the "outcome =" part outputs a linear model. Each number is an estimate of the β of the independent variable. That is, the numbers represent the magnitude of the impact of the independent variable on the estimate of the VP. Then, a linear model is constructed for each determined rule. In this study, two linear models were constructed.

Rule1 (60cases, mean −0.2952706, range −2.129606 to 2.573819)
　if
　　Leaning proficiency ≤ 0.8684222
　then
　　outcome (VP) = −0.0473442 + 0.42 VOIM + 0.19 IVPC + 0.02 Leaning proficiency

Rule2 (12 cases, mean 1.4763529, range 0.6924487 to 2.573819)
　if
　　Leaning proficiency > 0.8684222
　then
　　outcome (VP) = 0.417739 + 1.64 Effectuation − 1.04 Pre-commitments
　　　　−0.78 Causation + 0.74 Leaning proficiency

Evaluation on training data (72 cases):
　　Correlation coefficient 0.57

Fig. 2　Decision rule output using the Cubist function in R

To evaluate the performance of this model on the test data ($n=18$), the predict function in R was used to generate the predictions as in the regression tree. The correlation between the predicted values and the correct answers was also calculated using the cor function as in the regression tree.

The correlation between the prediction and the correct answer on the test data ($n=18$) using the model tree was 0.597, which is a considerable improvement from the correlation value for the regression tree, which was 0.260. Furthermore, the MAE is 0.591, which is lower than the MAE=0.767 of the regression tree. Thus, the model tree significantly improves the performance of the model as compared to the regression tree.

6. Discussion

The advantages of multiple regression analysis may be that (a) it is the most commonly used method for modeling numerical data and can be adapted to almost any data and (b) it expresses the magnitude and strength of the relationship between the independent and dependent variables in a concise linear relationship. However, its comes with some disadvantages: (a) the need to make assumptions about the data and the shape of the model in advance, (b) if there are many independent variables, a large number of models are created, resulting in a time-consuming process to extract the independent variables from them and select an appropriate model, and (c) in some cases, it is necessary to consider interactions where the independent variables have multiple influences, which increases the complexity.

Nevertheless, regression trees and model trees, which are decision trees that can predict numerical data, have the following advantages and disadvantages. The advantages are (a) the regression tree and the model tree combine both the data classification features of the decision tree and the ability to model numerical data, (b) there is no need to assume a model in advance, and (c) the decision tree classifies the selection of features (independent variables). Therefore, the decision tree is easy to use even when the number of features (independent variables) is large. As a result, (d) because the decision tree performs case classification, it may be better suited to real situations than multiple regression analysis. However, it also has the following disadvantages: (a) it is not as well known as multiple regression analysis, (b) it requires a large amount of training data, and (c) it may be more difficult to interpret than multiple regression analysis when the decision tree becomes large.

To clarify the factors and their relationship on ICV VP, this study first used multiple regression analysis. Although this approach did yield some results, the existence of complexity in terms of variable selection of independent variables to create an appropriate

model became clear, which is a drawback of the abovementioned multiple regression analysis. Although the three models selected using the stepwise method are statistically significant and possess the merit of explaining complex real-world problems with concise linear models, the question remains as to whether they are extremely simple and concise and thus deviate from the real world situation.

Therefore, this study tried to overcome the above shortcomings of multiple regression analysis by using a machine learning-based regression tree and model tree as a method to identify the factors affecting the VP of ICV.

First, the data was analyzed using the regression tree, which had the following three advantages: (a) the regression tree combines both the data classification features of the decision tree and the ability to model numerical data, (b) there is no need to assume a model in advance, and (c) the decision tree classifies the selection of features (independent variables). Hence, it is easy to use even when there are several features (independent variables). However, the accuracy of the prediction was not high.

Next, analysis was conducted using a model tree, which retains the three advantages of the decision tree described above while obtaining a relatively accurate prediction. In addition, while multiple regression analysis requires the selection of a single model, the model tree utilizes the decision tree, defining a series of decisions to classify the data and constructing a linear regression model for each classified rule. Thus, linear regression models were constructed for each classified rule similar to a flowchart, allowing the derivation of a decision model close to the actual situation. This study constructed two models affecting the VP of ICV; one for the case where the learning proficiency was less than or equal to 0.8684222 and the other for the case where the learning proficiency was greater than 0.8684222. When the learning proficiency was less than or equal to 0.8684222, the VP was 0.42, 0.19, and 0.02 higher in the cases of VOIM, where the parent company takes the lead in finding business opportunities; IVPC, where the value proposition at the time of ICV establishment is clear; and where the learning proficiency increases by one unit, respectively. In addition, when greater than 0.8684222, the learning proficiency was fully demonstrated. Therefore, effectuation, which consists of experimentation, affordable loss, and flexibility, and learning proficiency have a positive effect on VP, with values of 1.64 and 0.74, respectively. Nonetheless, causation, which is effective when the future is predictable, the goal is clear, and pre-commitment exists with customers, suppliers, and several other parties, had a negative effect on the VP: −0.78 and −1.04, respectively. It is satisfactory that the model was constructed to classify the cases in which the learning proficiency was greater than 0.8684222, which meant that the learning proficiency was sufficiently demonstrated, and less than or equal to 0.8684222, which meant that the learning proficiency

was not sufficiently demonstrated, and to express the relationship with VP separately.

The first contribution of this study is that, while it is important to identify the factors and relationships that affect the VP of ICV, the study tried to overcome the shortcomings of multiple regression analysis by using regression tree and model tree in combination with machine learning. In addition, as the model tree takes advantage of the decision tree's ability to classify data, a linear regression model could be constructed for each classified rule. In this way, each decision model is constructed based on the classified rule, rendering it possible to derive a decision model close to the actual situation, which is another contribution of this study.

7. Conclusion

Considering the importance of clarifying the factors and relationships affecting the VP of ICV, this study explored two new methods of machine learning in addition to the commonly used multiple regression analysis: One is regression tree, and the other is model tree. The regression tree predicts numerical values using the average value of the instances that reach a leaf node. The model tree is a hybrid model that combines the advantages of linear regression and decision tree, and constructs a regression model at each leaf node.

This study first attempted to clarify the factors and relationships affecting the VP of the ICV using a multiple regression model. Although this approach did yield some results, the shortcomings of complexity in terms of variable selection for independent variables in creating an appropriate model were still evident. In addition, although the linear regression model can explain complex real-world problems with a simple linear model, the possibility of deviations from the real world due to its simplicity and conciseness still remains.

Next, regression tree analysis was conducted, whose results showed that it was possible to utilize the benefits of the decision tree (it can both classify and model numerical data, does not require the assumption of the model in advance, and is easy to use even when the number of features (independent variables) is large because the decision tree classifies the choice of features (independent variables)). However, the prediction accuracy was not significantly high.

Finally, when analyzing the results using a model tree, it was possible to utilize the three advantages of the decision tree while providing relatively accurate estimation. In addition, because each decision model was constructed based on the classified rule, the model tree can derive a decision model close to the actual situation.

Machine learning, however, requires a large amount of training data. This study only used 72 training data samples. Therefore, it is necessary to increase the number of samples

and conduct analysis with sufficient training to improve the performance of the model. In addition, in this study, the regression tree and model tree used relatively simple learning methods. Therefore, it is critical to improve the accuracy of numerical prediction using more complicated machine learning methods in the future.

Acknowledgments

We would like to express our sincere gratitude to those who agreed to participate in this study and responded in the online questionnaire survey. We would also like to express our gratitude to Kunihiro Wakita for his help in conducting the questionnaire survey and compiling the data.

We would also like to thank Editage (www.editage.com) for providing English language editing services and publication support.

References

Chandler, G. N., DeTienne, D., McKelvie, A., & Mumford, A. (2011). Causation and effectuation processes: A validation study. Journal of Business Venturing, 26, 375–390

Covin, J. G, Garrett R. P. Jr., Kuratko D. F., Shepherd D. A., (2015). Value proposition evolution and the performance of internal corporate ventures. Journal of Business Venturing, 30, 749–774

Covin, J. G., Garrett, R. P., Gupta, J. P., Kuratko, D. F., & Shepherd, D. A. (2018). The interdependence of planning and learning among internal corporate ventures. Entrepreneurship Theory & Practice, 42(4), 537–570

Covin, J. G., Garrett, R. P., Kuratko, D. F., Bolinger, M. (2019). Internal corporate venture planning autonomy, strategic evolution, and venture performance. Small Business Economics, 56, 293–310

Garrett, R. P., Jr. and Covin, J. G. (2015) "Internal Corporate Venture Operations Independence and Performance: a Knowledge -Based Perspective". *Entrepreneurship Theory and Practice*. 39 (4). 763–790

Garvin, D. A. (2004). What every CEO should know about creating new businesses. *Harvard Business Review*, 82(7/8), 18–21

Johnson, K. L. (2012). The role of structural and planning autonomy in the performance of internal corporate ventures. *Journal of Small Business Management*. 50(3). 469–497

Kuratko D. F., Covin J. G. & Garrett R. P. (2009) Corporate venturing: insights from actual performance. Business Horizons, 52, 459–467

Lantz, B. (2019). Machine learning with R: Expert techniques for predictive modeling (3rd ed.). Packt Publishing

Sarasvathy, S. D. (2001) Causation and effectuation: Towards a theoretical shift from economic inevitability to entrepreneurial contingency. *Academy of Management Re- view*. 26(2). 243–263.

243–263

Sarasvathy, S. D. (2008). Effectuation: elements of entrepreneurial expertise. Northhampton, MA and Cheltenham: Edward Elgar.

Smolka, K. M., Verheul, I., Burmeister-Lamp, K., & Heugens, P. P. (2016). Get it together! Synergistic effects of causal and effectual decision makinglogics on venture performance. Entrepreneurship Theory and Practice. 42(4). 571–604

Quinlan, J. R. (1992) Learning with continuous classes. *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*. 343–348

Quinlan, J. R. (1993) Combining Instance-Based and Model-Based Learning. *In Proceedings of the Tenth International Conference on Machine Learning*. 236–243 University of Massachusetts. Amherst. Morgan Kaufmann