International Journal of Educational Media and Technology
2021, Vol. 15, No. 2 pp. 32-42

Risk Factors for Dropouts of University Students Assessed by Statistical Modeling

Naoto Yamashita
Media Opus Plus Inc., Japan
yamashita@mediaopusplus.com

Takato Takemori
Media Opus Plus Inc., Japan
taka-takemori@mediaopusplus.com

Hiroki Murakawa
Nihon Fukushi University., Japan
murakawa@n-fukushi.ac.jp

Osamu Sasagawa
Nihon Fukushi University., Japan
sasagawa@n-fukushi.ac.jp

Shingo Marubayashi
Media Opus Plus Inc., Japan
marubayashi@mediaopusplus.com

Shinichi Sato
Nihon Fukushi University., Japan
satoshin@n-fukushi.ac.jp

Shinji Nakamura
Nihon Fukushi University., Japan
shinji@n-fukushi.ac.jp

Supporting university students retain their studies is important for both the students and universities involved from educational and university management perspectives. Herein, we applied statistical modeling to questionnaire data regarding the dropouts of students upon enrolling in a private university of social welfare. Moreover, we considered the mechanism by which the students suspended their studies. The model used herein simultaneously expressed common trends across departments and department-specific mechanisms related to dropouts; it also predicted the probability of dropouts based on the responses to the questionnaire. Results suggested that the factors leading to dropouts varied across the departments in accordance with its specialties and whether the students took a national examination or not. The results were validated by the interviews of two experts, and they confirmed that the outputs of the proposed model were reasonable, while there still remained some important issues which seemed to be worthy of consideration in our future investigations.

Keywords: Dropout Prediction, Logistic Regression Analysis, Principal Component Analysis, Statistical Modeling

Introduction

According to a survey conducted by the Ministry of Education, Culture, Sports, Science, and Technology (2014) regarding the dropouts of university students, the percentage of students quitting their university is approximately 2.65% in Japan. Iwasaki et al. (2016) investigated the causes of student dropouts and their consequences. They pointed out that dropouts can be a cause of irregular employment status for students and negatively affect university revenues; therefore, dropouts result in a severe loss for our society. Yamamoto (2013) stated that dropout factors are complex and intertwined with multiple causes, including the mismatch of the student needs and educational content provided by universities as well as career anxiety. A recent and comprehensive review of the studies involving the reasons of

student dropout can be found in Aina et al. (2021), and they discussed the economic and sociological factors are critical to predicting students' achievements. According to their studies, we considered that it would be beneficial to students and universities to understand the mechanism of student dropout and develop dropout-suppression policies based on evidence-based knowledge concerning the dropout mechanism. In regard with dropout suppression, Ecker-Lyster et al. (2016) surveyed the dropouts in high school in the U.S. and emphasized the importance of early warning system of students' dropout based on educational data.

Recently, several statistical or machine learning approaches have been proposed to predict student dropouts. For example, Toyokawa (2015) proposed a prediction model using student grades, courses undertaken, and information on their enrollment as independent variables and investigated whether the student dropped out as a dependent variable. Okochi and Yamanaka (2016) and Kondo and Hatanaka (2016) attempted to predict dropouts based on machine learning models, such as gradient boosting decision trees and deep neural networks. Further, Takamatsu et al. (2019) proposed a model for dropout prediction, and importantly, their approach did not consider the interpretability of the model. For more examples, see Burgos et al. (2018) and Vihavainen et al. (2013). A systematic review by Agrusti et al. (2019) classified the 241 studies on dropout prediction according to their methodologies. Further, Kumar et al. (2017) surveyed the studies on dropout prediction in the universities in India.

However, because these studies proposed models with comparatively high generalization performances, they have the following two issues in effectively suppressing student dropouts. First, machine learning algorithms including the ones used in the aforementioned studies tend to be black boxes, and establishing a relation between independent variables and dropouts is difficult. In other words, understanding the dropout mechanism is difficult. It should be very important to share knowledge concerning dropout mechanisms among stakeholders and demonstrate that the policy is effective to make a dropout-suppression policy be effective. However, machine learning models are difficult to provide this explanation because they do not provide any suggestions regarding the mechanism of dropout. Second, the previous studies ignored the mismatch between student needs and educational contents provided by university, and they only focused on academic grades and achievements at their universities, instead of variables based on psychological tendencies of students. Recently, many universities are trying to develop their Institutional Research (IR) activities, and collecting a various data considered to be effective in supporting continuous and effective learning of their students. The relationships between dropouts and the following psychological aspects were suggested by some studies: psychological images of the university (Nakamura and Hamada 2015), expectations (Chishima 2018) of the university before and after the admission, Egograms (Goto et al. 2020), and University Personality Inventory (Koizumi 2017). Importantly, O'Neill et al. (2011) pointed out that, by reviewing the studies on dropouts in medical education, the effects of psychological factors on dropout were not well considered. To the best of our knowledge, however, there have been no studies that have attempted to construct prediction models employing these psychological aspects as independent variables. We hypothesized that the dissociation between student's expectations and reality recognitions against their university is a major factor of the dropouts, and it could be modulated by student's psychological tendencies.

Therefore, this study tried to establish a new dropout prediction model as a solution of the abovementioned issues. The proposed model is unique in the following point; instead of using a machine learning model, which tends to be a black box, it uses an interpretable model with understandable prediction logic. To achieve this purpose, we developed the logistic-RRMGR (L-RRMGR) model as an extension to the reduced rank multigroup regression (RRMGR) model proposed by Mayekawa and Yamashita (2020). In the L-RRMGR model, as is described hereinafter, it is possible to simultaneously expresses factors related to dropout mechanism both for sharing all departments in a university and specific to each of them. Moreover, the proposed model employed the independent variables that focused on students' expectations for their university and psychological characteristics. Herein, we picked up a case of a middle size private university mainly dedicated for education in a field of social welfare which is located in the middle part of Japan. Hereafter, we referred this university as University A. Specifically, we considered the student dropouts occurring in University A within four years as the signals to be predicted in the model, with employing the items obtained from student questionnaire concerning expectations for University A and psychological tendencies as the independent variables. This enabled us to focus on more fundamental factors than those considered in the previous studies and to achieve a prediction model with higher interpretability, which facilitates the understanding of the dropout mechanism. Further, the understanding would facilitate to update the current educational methods and contents as a dropout-suppression initiative, which could be an important suggestion for educational media and technology.

The current article reported methods to create the student dropout-prediction model and the results of the model applied to the dataset of the case of University A, with discussion of the dropout mechanism suggested by the estimated parameters. Furthermore, the suggested mechanism was validated by the interview sessions by educational experts. Finally, we tentatively proposed an effective dropout-suppression policy based on the outcomes of the model.

Method

Data

We obtained the questionnaire data from the students at University A in 2016 and 2017, and the subsequent dropout of these students. We excluded answers without student permissions to the use of the research. The 2016 and 2017 data were used as the training and validation data, respectively. Among the questionnaire items, the answers to the three questions X, Y, and Z (Tables 1–3) were used as independent variables. Because the answer format differed for each of the questions, we normalized values of each item ranging between 0 and 1.

Model

In this section, for introducing the model, we discuss some issues which should be considered. University A is composed of seven different departments. In a reality, University A has another department using a massive online education, but in this study, we focused on the departments with face-to-face education in person. We set up groups of the departments in terms of their specialties of education/learning, because there is a huge difference of the number of the students, and it may cause some difficulties in creating the prediction model. There are four department groups in accordance with their characteristics of the education (e.g., the students are supposed to take a national examination on their graduation, and so on). The departments groups are "welfare", "nursing/care", "education/psychology" and other "social sciences". Empirically, we know well that there is a great disparity in the motivation of studying and the reason of choosing the university between these department groups. It should be noted that the grouping we conducted here is considered to be appropriate since the educational curriculum, obtainable license, and the students' motivation are consistent within each of the four groups, and thus the grouping is valid only for University A.

Hereafter, we refer the integrated group of departments as "departments" for simplicity. Considering the disparity, it is possible to apply a single discriminant analysis model for all departments. However, parallelly applying a different model for each department would significantly increases the number of parameters to be interpreted and make it more difficult to understand both department specific and common factors related to dropout; thus, it is also considered inappropriate for the current purpose. Conversely, the factors common to all departments would provide important suggestions when promoting cross-disciplinary action in university for dropout suppression.

Mayekawa and Yamashita (2020) proposed a model called RRMGR model combining a principal component analysis, wherein many variables were compressed to a small number of components, and a regression model for multiple groups (Novick et al. 1972). This model, while extracting a small number of components common to all groups, expresses the impact (not always causal relationships) that these components have on dependent variables in terms of the regression coefficient for each group. Herein, as shown in the concept diagram in Figure 1, we developed an L-RRMGR model wherein the RRMGR model is modified for a dropout/non-dropout binary classification problem.

Table 1
Elements for which the absolute values of the estimated load matrix **A** *and the regression coefficient matrix* **W** *for question X. The elements in* **A** *that are 0.3 or above, or those in* **W** *that are 0.6 or above in absolute are displayed in bold type.*

	The questions below cover various aspects related to your life. For each question, please select one answer that most closely expresses how you feel.	Comp. 1	Comp. 2	Comp. 3
loading matrix **A**	Do you have the feeling that you don't really care about what goes on around you?	0.21	0.24	**−0.49**
	Has it happened in the past that you were surprised by the behavior of people whom you thought you knew well?	−0.01	**0.65**	0.29
	Has it happened that people whom you counted on disappointed you?	−0.10	0.12	−0.07
	Until now your life has had: no clear goals or purpose at all – very clear goals and purpose	−0.18	0.02	0.17
	Do you have the feeling that you're being treated unfairly?	0.12	**0.30**	−0.07
	Do you have the feeling that you are in an unfamiliar situation and don't know what to do?	0.15	**−0.36**	0.27
	Doing the things you do every day is: a source of deep pleasure and satisfaction – a source of pain and boredom	**−0.35**	−0.21	0.07
	Do you have very mixed-up feeling and ideas?	0.13	−0.23	−0.21
	Does it happen that you have feelings inside you would rather not feel?	**−0.37**	0.27	−0.01
	Many people – even those with strong character – sometimes feel like sad losers in a certain situation. How often have you felt this way in the past?	−0.13	−0.07	−0.13
	When something has happened have you generally found that: you overestimated or underestimated its importance – you saw things in the right proportion	**0.45**	0.10	**−0.42**
	How often do you have the feeling that there's little meaning in the things you do in your daily life?	**0.58**	0.09	**0.56**
	How often do you have the feeling that you're not sure you can keep under control?	−0.20	0.29	0.01
regression coefficient matrix **W**	Welfare department	−0.07	−0.38	−0.13
	Social sciences department	**−0.76**	−0.22	**0.60**
	Nursing/care department	0.05	−0.48	0.06
	Education/psychology department	**−0.65**	−0.37	−0.39

By setting the abovementioned departments as the groups of individuals assumed in this model, two types of tendencies in the dropout mechanism can be expressed simultaneously: the tendencies common to all departments and those specific to individual departments. The L-RRMGR model can be formulated by maximizing the following log-likelihood function:

$$LL(\mathbf{A},\mathbf{W}) = \sum_{g}^{G}\sum_{i\in\mathbb{G}_g}[\, y_i \mathrm{logit}(\mathbf{x}_{(i)}\mathbf{A}\mathbf{w}'_{(g)}) + (1-y_i)\{1-\mathrm{logit}(\mathbf{x}_{(i)}\mathbf{A}\mathbf{w}'_{(g)})\}\,] - \lambda||\mathbf{W}||^2 \quad (1).$$

Here, **X** expresses the N (students) × P (variables) matrix of independent variables, where $\mathbf{x}_{(i)}$ denotes the i-th row vector of **X** expressing the i-th student questionnaire response (i = 1, …, N), and y_i denotes 0 (non-dropout) or 1 (dropout) depending on whether the same student dropped out. Each student belongs to one of the G number of groups, and the set of students belonging to the g-th group (g = 1, …, G) is expressed as $\mathbb{G}_g$. The log-likelihood function is maximized with respect to the $P \times r$ (components) component loading matrix **A**, and the regression

Table 2
Elements for which the absolute values of the estimated load matrix **A** *and the regression coefficient matrix* **W** *for question Y. The elements in* **A** *that are 0.3 or above, or those in* **W** *that are 0.6 or above in absolute are displayed in bold type.*

	For the following statements, please select the level for each item that you feel most closely relates to you?	Comp. 1	Comp. 2	Comp. 3
loading matrix **A**	I am able to make an effective presentation using a variety of tools.	−0.10	**0.46**	−0.14
	I know about the "spirit of foundation" of the university.	0.08	**−0.34**	0.01
	I feel that I am a member of the "local area (place where I am living, place where I am studying)."	0.22	0.11	**−0.44**
	I am considerate of the physical and mental health of the people around me.	−0.04	0.07	**0.46**
	I understand the meaning of studying at the university.	**0.39**	−0.06	−0.19
	I know about the problems that adolescent university students are facing.	−0.16	−0.07	0.04
	I can understand the situation and feelings of the people subject to welfare.	0.02	0.14	0.12
	I am able to collect, organize, and process the necessary information on my own.	**−0.42**	−0.01	−0.12
	I am able to imagine or understand how other people see things around them.	**0.33**	0.03	−0.10
	I am able to make coherent assertions.	−0.19	**−0.44**	−0.07
	I am able to consider my own course of action at the university, linking my studies and extracurricular activities.	−0.07	−0.14	0.05
	I am able to plan and coordinate field work myself.	−0.13	0.23	0.18
	I am able to take leadership myself when performing group activities.	**0.50**	0.16	0.27
	I am able to appropriately summarize what people say and write.	0.29	−0.24	0.12
	I understand and am able to conduct the initiatives taken by my seniors at the university.	−0.09	0.15	**−0.44**
	I know what I need to do to live independently within the society.	−0.10	−0.26	−0.16
	I am able to think about the happiness of others.	0.25	−0.16	−0.19
	I am able to read the opinions and thoughts of others from their gestures.	0.01	0.01	**−0.33**
	I know what I need to do to be able to realize my future goals.	0.02	**0.41**	−0.05
regression coefficient matrix **W**	Welfare department	0.33	−0.38	**1.06**
	Social sciences department	−0.16	**1.24**	0.59
	Nursing/care department	**−0.87**	−0.35	0.12
	Education/psychology department	**−1.23**	−0.33	−0.07

coefficient matrix **W** of $G \times r$. $n^{-1}\mathbf{A}'\mathbf{A} = \mathbf{I}_r$ is imposed on **A** to exclude the indeterminacy of transformation using an arbitral nonsingular matrix, where $\mathbf{I}_r$ represents an r-dimensional identity matrix. Here r ($\leqq \min(P, G)$) expresses the number of components which is fixed beforehand. The dropout probability of student i having the feature vector $\mathbf{x}_{(i)}$ is obtained as $\mathrm{logit}(\mathbf{x}_{(i)}\mathbf{A}\mathbf{w}_{(g)}')$, using the g-th vector of **W** noted as $\mathbf{w}_{(g)}$ corresponding to the g-th group to which the student belongs, the matrix **A**, and the logit function $\mathrm{logit}(x) = 1/\{1 + \exp(-x)\}$. The novelty of the proposed model is that the operation refers the component load matrix **A** which is common for all groups and the regression coefficient matrix $\mathbf{w}_{(g)}$ specific to the group g. Furthermore, because the logit function is bounded as [0, 1], $\mathrm{logit}(\mathbf{x}_{(i)}\mathbf{A}\mathbf{w}_{(g)}')$ can be interpreted as the probability that student i belonging to group g will dropout from university. However, to prevent the overfitting of the training data, we employ the L_2 regularization with the regularization parameter λ (> 0). Hereinafter, $\lambda = 1.0$ is used for simplicity.

In the L-RRMGR model, even if **A** is converted to **AT** and **W** to **WT** using an r-dimensional orthonormal matrix **T** satisfying $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_r$, the value of the log-likelihood function does not change before and after convergence,

Table 3
Elements for which the absolute values of the estimated load matrix **A** *and the regression coefficient matrix* **W** *for question Z. The elements in* **A** *that are 0.3 or above, or those in* **W** *that are 0.6 or above in absolute are displayed in bold type.*

	For the following statements about the motivation of your study at the university, please select the level for each item how clearly manifests your motivation.	Comp. 1	Comp. 2	Comp. 3
loading matrix **A**	Because I am curious about it	−0.24	**0.41**	0.01
	Because I was urged to do so by those around me	0.01	0.06	**0.50**
	Because it is linked to my future success	**−0.36**	0.02	0.00
	Because I want to get good grades and evaluations.	0.14	**0.59**	0.01
	Because it is necessary for the employment exam and work	−0.16	**−0.58**	0.13
	Because I would be anxious if I did not	0.05	0.01	**0.64**
	Because the teaching materials and books are interesting	**0.64**	0.06	0.14
	Because people around me would complain if I did not	−0.11	0.24	−0.36
	Because I do not want to fall behind from the other students around me	**0.37**	−0.11	−0.18
	Because I am happy to be able to understand the content	**−0.35**	−0.07	−0.07
	Because it will be useful for various things in the future	−0.29	0.21	**0.36**
	Because my parents told me to do so	−0.03	0.15	0.07
regression coefficient matrix **W**	Welfare department	−0.39	−0.28	**−0.85**
	Social sciences department	−0.45	0.32	−0.42
	Nursing/care department	**1.02**	0.23	0.05
	Education/psychology department	−0.01	**0.96**	−0.02

shown as $\mathrm{logit}(\mathbf{x}_{(i)}\mathbf{A}\mathbf{w}_{(j)}') = \mathrm{logit}(\mathbf{x}_{(i)}\mathbf{A}\mathbf{T}\mathbf{T}'\mathbf{w}_{(j)}')$. Therefore, by rotating **A** using different orthogonal rotation methods, obtaining an easily interpretable **A** is possible. In this study, after estimating the parameter matrix, **A** was rotated using the Varimax rotation (Kaiser 1958).

For the parameter estimation, the log-likelihood function is optimized through numerical optimization using the L-BFGS-B method. To avoid localized solutions, we started the optimization from 100 initial values and accepted the group of solutions for which the function value was the largest as the final solution. The number of components was set to three for interpretability.

Evaluation

To evaluate the performance of the trained models, we conducted quantitative and qualitative evaluation. In quantitative evaluation, some metrics that measure how accurate the model predicts the student dropouts were computed, with respect to the training data and the validation data, respectively. This evaluation serves to quantify the generalization performance of the trained model. Subsequently, we interpreted the estimated parameters of the trained model, and named the department-specific risk factors for student dropouts and built the hypothetical mechanism of the dropouts. Two student-support experts qualitatively evaluated these risk factors and dropout mechanism.

Results and Discussion

Prediction Performance

To evaluate the performance of the three trained models, we computed the ROCAUC values for the training and validation data obtained in 2016 and 2017. ROCAUC ranges in [0, 1], and its maximum value 1 represents a high discrimination ability. As shown in Table 4, ROCAUCs of approximately 0.7 were obtained for the training data, confirming that the model was well trained. However, the validation data values were slightly lower at approximately 0.6, indicating that the prediction performance was lower than that for the training data.

Moreover, Table 4 also showed that the recall (ratio of students that could be predicted to dropout from among all dropouts) was about 0.7 for the training data, with the threshold of dropout probability was set to 0.1. This implies that 70% of the students who actually dropped out could have been predicted. Furthermore, the precision (ratio of

students predicted to dropout that actually dropped out) ranged between 0.177 and 0.099 both for the training and validation data. Thus, it should be concluded that it would be difficult to predict 2017's dropouts by the model trained by 2016 data. Therefore, we focused on understanding the mechanism of students' dropouts by interpreting

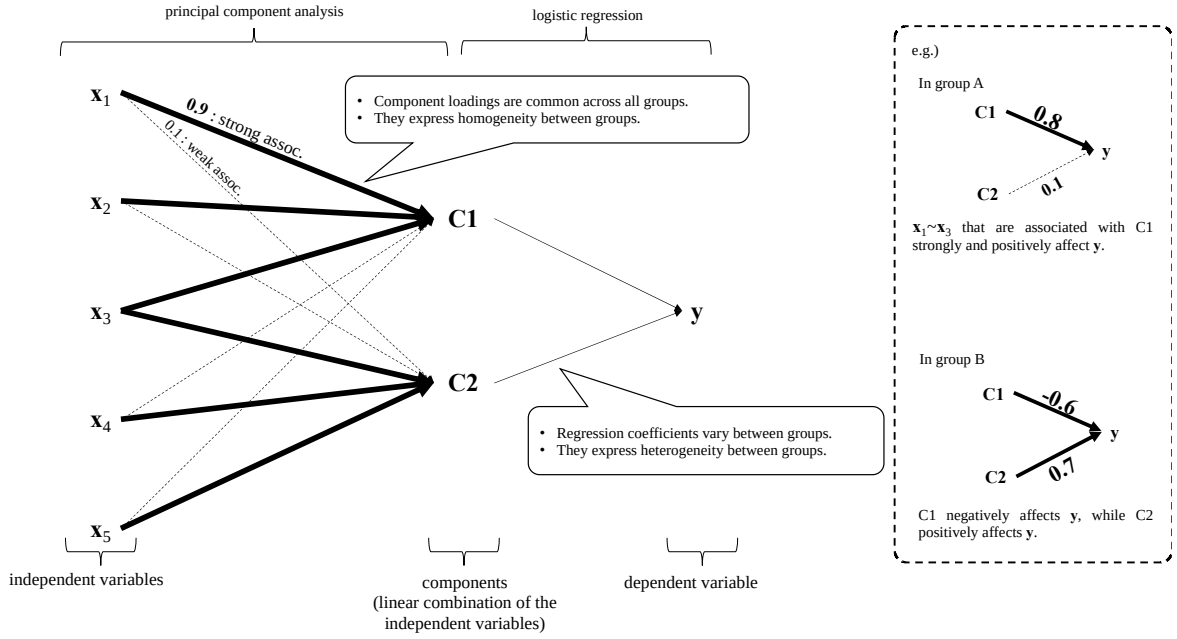

Figure 1. Concept diagram of the logistic-RRMGR model.

the estimated parameters, that would be beneficial in an attempting to make a policy that deterrence student's dropout. It is only capable with our interpretable model, but incapable with a conventional black-box-based machine learning model.

Interpreting Common Components among Departments

First, we shall interpret the components expressed in the matrix **A**, which expresses the factors related to dropouts common to all departments. In Tables 1–3, the Varimax-rotated loading matrices are shown for each of the three questions. Focusing on the questionnaire items with comparatively high loadings, the three components obtained from question X can be interpreted as "meaningfulness," "understandability," and "autonomy." The format of question X is known as a sense of coherent scale (Antonovsky, 1987). For its interpretation, we referenced the work of Togari and Yamazaki (2005) on factor structures.

Similarly, for question Y, which asked the skill set when entering the university, we obtained components interpreted as "communication-related skills," "high level of goal awareness and lack of significance in studying at University A," which positively loads on "I know what I need to do to be able to realize my future goals" but negatively loads on "I know about the "spirit of foundation" of the university" in question Y, and "lack of cooperativeness with those around." Finally, for question Z, which asked motives for learning, we obtained components interpreted as "interest-oriented," "necessity-oriented," and "extrinsically motivated."

Department-specific dropout risk factors

The lower sections of Tables 1–3 indicated how the common components to all departments interpreted above were to promote or suppress dropouts in each department. This can be understood by referring to the regression coefficient matrix **W**. For example, the first component of question X interpreted as "meaningfulness" exhibited a high negative regression coefficient for the social sciences and education/psychology departments. This indicates that students who feel "meaningfulness" in these departments are unlikely to dropout. Conversely, the students with low "meaningfulness" were supposed to be a subject of a higher dropout risk. The dropout risks for each of the departments as follows.

First, for the welfare departments, "lack of cooperativeness with those around" (the third component of question Y)

exhibited a high positive regression coefficient, while "extrinsically motivated" (the third component of question Z) turned out to have negative coefficient with significant higher absolute value. Thus, in these departments, anxiety regarding their skills for cooperating with others at the time of entrance to the university and low extrinsic motivation for studying would increase the risks of dropout. For the former factor, many students belonging to the departments have a motivation to engage in a field of social welfare where the close communication with colleagues and care-receiver should be necessary, and thus, "lack of cooperativeness" would act a facilitative
factor for the dropouts in this departments. The students with higher anxiety to cooperate with others seemed to also have higher anxiety for their future job in this departments, resulting in their higher risk of the dropouts.

Table 4
ROCAUC, precision, and recall values when the threshold is 0.1 for the training and validation data and the questions X, Y, and Z.

	ROCAUC			Precision			Recall		
	X	Y	Z	X	Y	Z	X	Y	Z
Training data	0.720	0.711	0.712	0.177	0.171	0.170	0.696	0.696	0.705
Validation data	0.568	0.565	0.527	0.114	0.107	0.099	0.510	0.500	0.469

Subsequently, in the departments of social sciences, "high level of goal awareness and lack of significance in studying at University A," low "meaningfulness," and high "autonomy" were considered to act as dropout risks. Because these departments are less related with a field of social welfare, which is a primary advantage of University A, and thus, quite unique as compared with the other departments, the students might feel that there is no significance to studying at a university dedicated for an education in the field of social welfare. The positive correlation between the dropouts and less significance of studying a University A might suggest that a quite a few students in these departments have a doubt in studying at the university characterized as a national center of the research/education in the field of social welfare. Moreover, it is indicated that the students with stronger autonomy and goal awareness turned out to have a higher risk of the dropouts, suggesting that the students in this departments who find their life objectives in other than learning at the university might consider changing their life courses. It may also be consistently explained by the specialty of these departments.

For the nursing/care departments, low "communication-related skills", and high "interest-oriented" motivation for study, rather than "necessity-oriented" motivation, in terms of learning motives can be risk factors for dropout. All students with very few exceptions in these departments will take national examinations to get their professional qualifications, at the time of their graduation. Most of the courses in the departments are designed for preparation of the exams. The result suggested that the "necessity-oriented," not the "interest-oriented," motivation would have an advantage in the retention under such situation. It would be reasonable to assume that deeper intrinsic motivations, i.e., interest-oriented, would be more effective in a continuous learning. The current result was not consistent along with this line. The efficacy of learning motivations may depend on a situation where the student learns and a future goal what the student wants to get.

The factors related to dropouts in the education/psychology departments were low "communication-related skills" and low "meaningfulness." The former shared the tendency with the departments of welfare and the latter was consistent with the departments of social sciences. In these departments, some students want to be professionals in educational or psychological fields after their graduation, while the others tend to have more general carrier design. Thus, it is no so surprising that the result for the education/psychology departments exhibited somehow intermediate mixtures between the departments of the welfare and the social sciences. Contrary to the nursing/care departments, higher "necessity-oriented" motivation turned out to be resulted in higher risk of dropouts in these departments. The contrast would be due to differences of the educational curriculum in the departments; education in the education/psychology departments are mainly designed based on liberal arts rather than developing specific professionals.

Interviews to the Experts

To validate the above discussions on the dropout risks in the four departments, we conducted interview sessions with two student-supports experts.

The sessions were organized as follows. After a brief introduction of the interviewees and interviewer and ice breaking talks, the interviewer introduced the summary of the research result, which focused on the department-specific risk factors for dropout discussed in the previous section. Subsequently, the interviewer asked the interviewee whether they could agree with these risk factors and the reasons for their agreement/objection. The interviewees were asked to evaluate our research findings qualitatively based on their student-support experiences. The whole session was approximately 45-min long.

In the interview session with the first interviewee, who had worked as a student adviser at University A for a year, the interviewee answered that he could agree with the risk factors found, especially with "lack of cooperativeness with those around" in the welfare departments. He pointed out that the students in these departments are required to get the degrees of childcare and nursing training, where cooperation with others is necessary, and "lack of cooperativeness with those around" would be a risk of quitting the training and subsequent dropout. Moreover, he agreed that a lack of "necessity-oriented" motivation for studying is a dropout risk in the nursing/care departments because their educational curriculum was designed to get the professional licenses of nursing or caregiving. The students must devote their best efforts to pass a national examination for the license, and thus clear and strong motivation for the study will be required. Importantly, he objected our model because it did not consider whether students had enough time for their study. In his experience, some students gave up their study for the national exam because they could not have enough time for study even if they were strongly motivated.

The second interviewee was an associate professor at University A and assigned as an adviser for students at the welfare departments. He pointed out, as well as the first interviewee, that "lack of cooperativeness with those around" is a dropout risk in the welfare departments. Further, he considered that strong "autonomy" and "high level of goal awareness and lack of significance in studying at University A" as risk factors for dropouts in the departments of social sciences. His consideration was because there was less difference of educational contents between University A and other neighboring universities in this field. Some students might have difficulties to find a significance in studying social sciences other than social welfare at University A, because the university has been well recognized as having advantages in education of welfare for long years.

The second interviewee also pointed out the limitations of our model. First, he mentioned that perceived affinity of a student toward the teaching staff should be added as an independent variable in the prediction model because several students have dropped out from their classes due to that they could not manage good relationship with the faculty members. Second, although we handled all the students in the education/psychology departments as same student group in our model, it is composed of two different courses: Psychology course and Education course actually. In the latter course, most of the students are interested in acquiring national licenses for primary and secondary school teachers, whereas the first course is oriented liberal arts; therefore, the motivational background of the students should not be considered as homogeneous. He thus suggested that the latter course should be merged to the nursing/care departments.

Based on the two interview sessions, we confirmed the overall validity of the dropout risk factors found in University A derived from the dropout prediction model. Two experts pointed out the limitations of the model, which should be considered in future studies.

Conclusive Remarks

In this study, we developed a new model for predicting and explaining student dropouts from their university, attempting to obtain valuable knowledge in determining student-support policy for encouraging retention and suppressing dropout. The estimated model parameters and their interpretation suggested the dropout mechanism and risk factors that are common/specific to the departments, in the latter of which would reflect the educational content and student career paths in each department. The findings were evaluated by two experts, and they equally pointed out that the findings were reasonable based on their experience.

Based on the current findings, we can propose tentative actions for dropout suppression as follows. First, "a lack of communication-related skills" was found to be a common dropout risk in the nursing/care and the education/psychology departments. Programs of basic communication skills applied to the first-year students in these departments would be effective. We also found that "meaningfulness" is a dropout risk factor in multiple departments. The "meaningfulness" is a one of the three subscales of "sense of coherence (Antonovsky,1987)," which has been considered as a strong predicter of one's behavior in many aspects (e.g., Togari and Yamazaki, 2005). Therefore, it should be effective to apply prior screening with the sense of coherent scale on university entrance and conduct a general follow-

up on students who feel low "meaningfulness". For the future study, the effectiveness of the above dropout-suppression policies should be investigated, and we believe the research would be an important suggestion for the future of educational methods.

The generalization performance of the proposed model proposed herein was not so high in terms of the ROCAUC value for the validation data. This is because we applied an interpretable model rather than a black-box model, such as the recent machine learning models. The L-RRMGR is thus designed to be a generalized linear model. In our future study, we will try to increase prediction performance by both improving the interpretable model and integrating a black-box oriented machine learning models, such as gradient boosting trees, followed by interpretation referring to feature importance.

It should be noted that the outputs from the model are not necessarily indicate causal relationship between dependent and independent variables, because the model was based on the logistic regression, although the causality was somehow validated by the expert interviews. Further research should be executed to confirm the causality between the risks and dropouts. For example, dropout-suppression policies should be tested in a group of randomly selected students, and the resulting dropouts should be compared with those without the treatment. These trials would be beneficial to accumulate deeper knowledge concerning student's dropout and to develop an actual and effective method to prevent it.

References

Agrusti, F., Bonavolontà, G., & Mezzini, M. (2019). University Dropout Prediction through Educational Data Mining Techniques: A Systematic Review. *Journal of E-Learning and Knowledge Society*, 15(3), 161-182.

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2021). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*.

Antonovsky, A. (1987). *Unraveling the mystery of health: How people manage stress and stay well*. Jossey-bass.

Burgos, C., Campanario, M., de la Peña, D., Lara, J., Lizcano, D., & Martínez, M. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers and Electrical Engineering*, *66*, 541-556.

Chishima, Y. (2018). *Do expectations of campus life predict adjustment to campus life? A one-year longitudinal study toward pre-university students*. 82nd Conference of Japanese Psychological Association (in Japanese).

Ecker-Lyster, M., & Niileksela, C. (2016). Keeping students on track to graduate: A synthesis of school dropout trends, prevention, and intervention initiatives. *Journal of At-Risk Issues*, 19(2), 24-31.

Goto, F., Hashimoto, T., Nakagawa, M., Okamoto, H., Sato, K., & Holsworth, M. J. (2020). Graduation Prediction of University Student Athletes by Egograms and Suggestions to Improve Academic Performance. Koto Kyouiku Forum (in Japanese).

Iwasaki, Y., Miyajima, K., Kagehisa, T., Fukushima, K., & Taninouchi, S. (2016). Discussion on the prevention of dropout. *Kochi University Reports on Rducational Research and Activity*, *20*. 49-60.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187-200.

Koizumi, S. (2017). Relations between UPI scores at entrance to university and taking temporary absence, and dropping out early. *Departmental Bulletin Paper of Kyoei University*, *15*, 73-92 (in Japanese).

Kondo, N., & Hatanaka, T. (2016). Modelling of Students' Learning States Using Big Data of Students through the Baccalaureate Degree Program. *Transactions of Japanese Society for Information and Systems in Education*, *33*(2), 94-103 (in Japanese).

Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on educational dropout prediction. *International Journal of Education and Management Engineering*, 7(2), 8.

Mayekawa, S., & Yamashita, N. (2020). Bayesian reduced rank multigroup regression analysis: A new model for multigroup data with hybrid parameter sharing. *Behaviormetrika*, 47(2), 411-426.

Ministry of Education, Culture, Sports, Science and Technology-Japan. (2014). *Situation regarding student mid-term dropout and leave* . Retrieved May 31, 2021 from https://www.mext.go.jp/b_menu/houdou/26/10/__icsFiles/afieldfile/2014/10/08/1352425_01.pdf (in Japanese)

Nakamura, M., & Matsuda, E. (2014). Impact of identification with the university on non-adaptability to the university-Analysis using attendance rate and GPA [Translated from Japanese]. *Bulletin of Edogawa University* (in Japanese).

Novick, M., Jackson, P., Thayer, D., & Cole, N. (1972). Estimating multiple regressions in *M* groups: A cross-validation study. *British Journal of Mathematical and Statistical Psychology*, 25(1), 33-50.

Okochi, Y., & Yamanaka, A. (2016). Early discovery of students with poor grades in early years, using data, such as placement test and high school course status [translated from Japanese]. *Japan Journal of Educational Technology*,

40(*1*), 45-55.

O'Neill, L. D., Wallstedt, B., Eika, B., & Hartvigsen, J. (2011). Factors associated with dropout in medical education: a literature review. *Medical Education*, 45(5), 440-454.

Togari, Y., Yamazaki, Y., Nakayama., K., Yokoyama, Y., Yonekura, Y., & Takeuchi, T. (2015). 13-item 7-case method-sense of coherence scale - Calculation of Japanese version criteria [translated from Japanese]. *Japanese Journal of Public Health*, *62*(5), 232-237 (in Japanese).

Toyokawa, K. (2015). A study of academic adviser support by analytics of student data. Studies in International Relations. *Institute for National Policy,* 36(1), 79-86 (in Japanese).

Vihavainen, A., Luukkainen, M., and Kurhila, J. (2013). Using students' programming behavior to predict success in an introductory mathematics course. In *Educational Data Mining 2013*.

Yamamoto, S. (2013). What is students' dropouts?: Their mechanism, reason, and challenge in their prevention [translated from Japanese]. *Kochi University Educational Journal*, *35*(1), 20-24 (in Japanese).