Predicting the Probability of Student Dropout through EMIR Using Data from Current and Graduate Students

Katsuhiko MURAKAMI
Human Genome Center,
The Institute of Medical Science,
The University of Tokyo,
Tokyo, Japan

○Kunihiko TAKAMATSU*
Faculty of Education,
Organization for the Advancement of Higher Education,
Center for the Promotion of Excellence in Research and Development of Higher Education,
Life Science Center
Kobe Tokiwa University,
Kobe, Japan
ktakamatu@gmail.com
*Corresponding Author

Yasuhiro KOZAKI
Home Economics Education,
Department of Education,
Osaka Kyoiku University,
Osaka, Japan

The Center for Early Childhood Development, Education, and Policy Research,
The University of Tokyo,

Aoi KISHIDA
Kobe City Nishi-Kobe Medical Center,
Kobe, Japan

Kenya BANNAKA
Department of Oral Health
Center for the Promotion of Excellence in Research and Development of Higher Education
Kobe Tokiwa College,

Ikuhiro NODA
Instruction Section
Organization for the Advancement of Higher Education,
Center for the Promotion of Excellence in Research and Development of Higher Education,
Kobe Tokiwa University,

Jyunichiro ASAHI
Head Office,
Kobe Tokiwa University
Kobe, Japan

Kazuyoshi TAKAO
Institutional Research Office,
Kobe Tokiwa University,
Kobe, Japan

Kenichiro MITSUNARI
Faculty of Education,
Academic Affairs Committee
Kobe Tokiwa University,
Kobe, Japan

Regional Liaison unit,
Center for the Promotion of Interdisciplinary Education and Research,
Kyoto University, Kyoto,

Tadashi NAKAMURA
Tamada Educational Institution
Kobe, Japan

Yasuo NAKATA
Faculty of Health Sciences,
Organization for the Advancement of Higher Education,
Kobe Tokiwa University,

***Abstract*—In 2016, an office for institutional research (IR) promotion was constructed at Kobe Tokiwa University. The purpose of this office is to propose, manage, arrange, and collect information on students at the university not only as a general management strategy, but also to provide enrollment management. Work at the IR promotion office officially began 2016, at which point we began to perform university data management and collection duties. A promotion unit for IR was also newly established in 2017. In contrast to IR promotion practices in which the department is run solely by an administrative staff, the IR promotion unit at Kobe Tokiwa University is a collaborative group consisting of both staff and faculty. In this unit, we carried out an analysis focusing on the "student dropout " by using data from both current, graduate, and dropout students.**

Our database currently contains 3,495 points of data (i.e., headcounts), each containing 1,246 items of numerical value. We obtained results with these data by processing them through a machine learning technique using three methods; a logistic regression yielded a test correction rate exceeding 95%, while a random forest yielded a correction rate of about 90%.

Keywords—dropout, machine learning

I. Introduction

A 2008 central council report titled "Toward building bachelor degree education" clarified that teaching management and professional faculty and staff development were subject to three policies (i.e., those involved in awarding degrees, curriculum implementation and organization, and student enrollment) [1]. Subsequent to this report, scientific thinking styles involving statistical data and "evidence based" research gradually became required at the university level.

Following this trend, regulations for the enforcement of school education laws were renewed in 2011. In addition, clarifications were made regarding the educational information that each university was required to provide. A "University Portrait" was then developed by the central council in 2012 as a glossary providing information "Toward the qualitative change of university education to build a new future: To the university that continues to learn life and develops the ability to think independently." Here, we found a description related to IR stating that "because of the need to maintain a university portrait (temporary name), [1] the university should grasp and analyze its activity by using educational information to lead reform (i.e., improve the function referred to as institutional research (IR))" [2]. Japanese interest in university IR increased after this report was published. A steering committee and university portrait center were then established at the National Institution for Academic Degrees and Quality Enhancement of Higher Education. University portraits were disclosed for private schools on October 6, 2014 [3], while those for public schools were published on March 10, 2015 [4].

We followed this trend at Kobe Tokiwa University by establishing an IR committee in preparation for the 2015 term before establishing an IR promotion office in 2016. At our university, IR consists of proposing, managing, arranging, and collecting information on students. This is not merely a general management strategy, but also one aimed at enrollment management. Operations at the IR promotion office initiated last year, at which point we began collecting and arranging data.

An IR promotion unit was established this year. In contrast to IR promotion practices in which the department is run solely by an administrative staff, the IR promotion unit at Kobe Tokiwa University is a collaborative group consisting of both staff and faculty. In this unit, we carried out an analysis focusing on the "dropout of students" [5] as part of our main inquiry.

Student dropout has been studied at the macro and micro levels. Macro-level study is based on data from basic survey reports conducted by the school [5], [6], while micro-level study involves the student counseling approach, which is aimed at preventing dropouts [7]. Regarding enrollment management institutional research (EMIR), dropout problems have been studied at both Yamagata University [8] and Kyoto Koka Women's University [9]. Dropout prevention has also been studied through Grants-in-Aid for Scientific Research with the goal of obtaining data through EMIR [10].

The term "dropout" is not only related to positive or spontaneous events such as "studying abroad" or gaining "admission to another university," but also has negative connotations including "economic difficulty" and a decrease in the desire to learn [5]. Since they occur for many reasons, it has been difficult to predict student dropouts. To confront this problem, we carried out an analysis focusing on the "dropout of students" through evidence-based research.

We successfully investigated the probability of student dropout through EMIR on graduate students in 2017 [11]. In this study, we added data from current students to predict dropout probability through three machine learning methods.

II. Methods

A. Data preparation

As a general rule, Kobe Tokiwa University's EMIR data is not publicly available. For this reason, data were given anonymously (i.e., all student ID numbers and personal identifiers were deleted at the IR promotion office, and all items were concealed by converting the numerical labels to protect the contents). Thus, individuals conducting the analysis were not able to determine the provenance of the data. However, the IR promotion office alerted us that the first items used in the machine learning process were data indicating student dropouts, and that data directly representing dropouts were excluded.

Data were obtained from the IR promotion office in comma-separated values format. The data were then saved in tab-separated values format before being converted into line feed code for Unix. At that time, the existing cells included double line feed, which we then modified before converting the Kanji characters to UTF-8 using nkf. Missing data were treated as 0.

B. The machine learning technique

Analyses were performed on Mac OS X 10.11.6 using Python (3.6.0) and Perl (5.18.2). For Python library, we used numPy [12], matplotlib [13], scikit-learn [14], and pandas [15]. For machine learning, we used 70% of the training set data and 30% of the test set data.

III. Results and Discussion

The IR promotion office provided us with points of data for 3,495 combined current and graduate students. A total of 1,246 items existed for each data point. The first of those items indicated the correct answer for machine learning (i.e., the data indicating student dropouts).

Our research question involving the "discussion of predicting the probability of student dropouts using a machine learning technique for EMIR" can be paraphrased as "Can binary discrimination be predicted regarding dropout or graduation using 1,246 values for each of 3,495 total data points from both current and graduate students through machine learning?"

The field of machine learning is associated with the "no free lunch" theorem [16]. This theorem indicates that there are events that should be given attention during machine learning, meaning that there are no machine learning methods that can produce high precision data for every problem. With this in mind, we performed machine learning using three methods. Logistic regressions were used as the first and second methods [17].

TABLE 1. CORRECT ANSWER RATE FOR THE TRAINING AND TEST SETS OF THE THREE MACHINE LEARNING METHODS.

Method Correct answer rate	1	2	3
Training accuracy	0.988	0.992	0.941
Test accuracy	0.969	0.963	0.929

A regression analysis is a method used to investigate the effects of a variable, which is predicted by formulas composed of objective variables (also called dependent variables; in this study, these variables are “dropout” and “graduate”) and explanatory variables (also called independent variables; in this study, these are the previously mentioned 1,246 items).

A logistic regression is a recurrent algorithm used to convert qualitative variables into linear form. In this study, the variables represented two values (distributive) that were regressed to linear form using a logistic regression. A usual regression returns mostly matched results to the training set data, with which the machine might be induced to overlearn. To avoid overlearning, we used the L1 and L2 regularization methods for the regularization term (also called the penalty term). The L1 regularization term uses the sum of absolute values of coefficients, while the L2 regularization term uses the sum of squared coefficients. The first method used L2 as regularization, while the second method uses L1 as regularization. We used a logistic regression featuring a solve issue that could not be linearly separated when the issue was judged to be greater than the cubic equation.

The third method we used was the random forest [18]. Random forest is a machine learning algorithm that uses ensemble learning. This algorithm employs multiple decision trees as weak classifiers and integrates the results (i.e., the forest) to gain correct results. Random forest featuring is available to use pattern recognition, regression, and clustering.

The results of these methods are shown in Table 1. In the first machine learning method, the correct answer rate for the training set was 0.643, and the correct answer rate for the test set was 0.649. We were not able to determine the difference between the correct answer of the test and the training. For the second machine learning method, the correct answer rate for the training set was 0.573, and correct answer rate of the test set was 0.603. For this method, the correct answer rate of the test set was higher than the correct answer rate of the training set. The difference between the first and second methods was the method of regularization. From these results, we determined that student dropout for the L2 regularization correct answer rate was higher than that of the L1 regularization rate.

For the results of the third machine learning method, the correct answer rate for the training set was 0.914, and the correct answer rate for the test set was 0.895. The correct answer rate was about 25 points higher for the random forest when compared to the logistic regression.

We then extracted the top 20 items that contributed to our prediction (Figures 1, 2 and 3). For the first and second methods, we output the coefficient for each of the items. For the third method, we output the importance of the property score. The value evaluation for the third method involved the numerical value obtained by normalizing the degree to which the accuracy decreased when that information disappeared. Thus, there is no good direction on the scale (i.e., whether the student is able to graduate) if the number is large, which shows the magnitude of its influence on judgement.

Fig. 1 and Fig. 2 indicate that item 72 was very important for determining student dropout. Unfortunately, we do not know the contents of item 72 because the IR team was not permitted to reveal them.

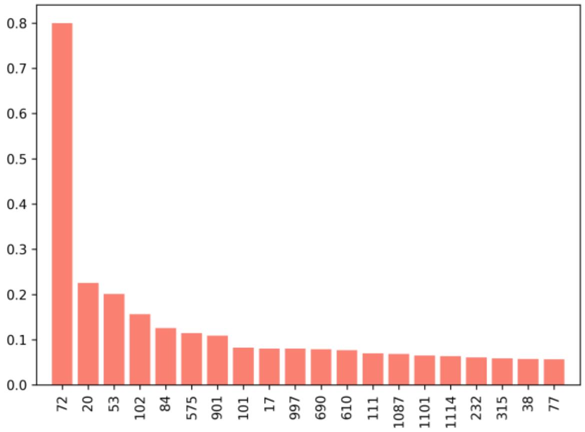

Fig. 1. The top 20 item numbers contributing to the predictions of the first method

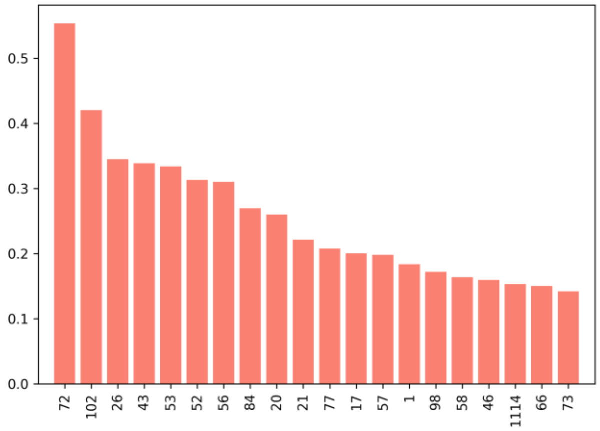

Fig. 2. The top 20 item numbers contributing to the predictions of the second method

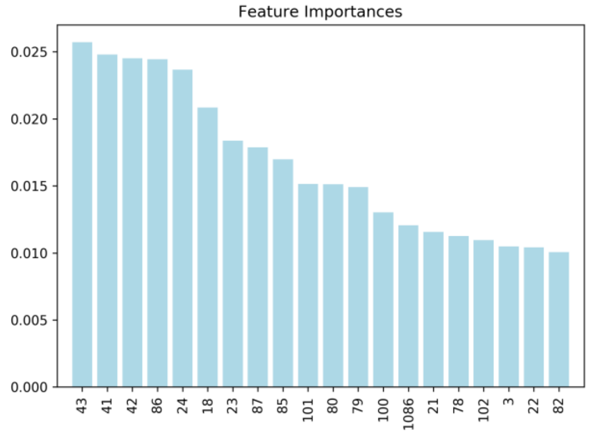

Fig. 3. The top 20 item numbers contributing to the predictions of the third method

Item 72 does not appear in third methods (Fig. 3). This means that the importance or contribution of this item was different between the three machine learning methods.

Our final study indicated that that test accuracy of the same methods when only examining data from graduate students was 0.649, 0.603, and 0.895, respectively. This means that the test is more accurate when using data from both current and graduate students than it is when only examining data from graduate students.

In this analysis, we performed machine learning to predict dropout rates by using the data of both current and graduate students. However, our primary issue of interest is dropout prevention. It is therefore important to continually use these methods to predict student dropout rates each week. Regarding EMIR, it is important to investigate unknown items when predicting these dropout rates through machine learning. Prior to our study, it was revealed the data used to directly indicate dropout rates were not included in the dataset with which we were provided. However, many items in the data we used seemed to indirectly indicate dropout rates. Thus, the IR promotion office should investigate the data used in our analysis when operations begin. We plan to promote this study as a capable method of predicting student dropout through machine learning.

REFERENCES

[1] MEXT, "Construction of education for university," *The Central Council for Education*, 2008. [Online]. Available: http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo4/houkoku/080410.htm. [Accessed: 15-Mar-2018].

[2] MEXT, "Towards the qualitative transformation of university education in order to build a new future," *The Central Council for Education*, 2012. [Online]. Available: http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/1325047.htm. [Accessed: 15-Mar-2018].

[3] PortrateCenter, "Portrate of univerisity for private univeristy," *National insitution for academic degrees and quality enhancement of Higher Education*, 2015. [Online]. Available: http://up-j.shigaku.go.jp/. [Accessed: 15-Mar-2018].

[4] PortrateCenter, "Portrate of univerisity for public univeristy," 2016. [Online]. Available: http://portraits.niad.ac.jp/. [Accessed: 15-Mar-2018].

[5] K. Anegawa, "The Impact of Learning and Living Environments of Colleges on Dropout Rates," *Kyushu Daigaku Daigakuin Keizaigakukai*, vol. 149, pp. 1–16, 2014.

[6] F. Maruyama, "A Study in Dropout in Japanese Colleges and Universties," *J. Educ. Sociol.*, vol. 39, pp. 140–153, 1984.

[7] S. Kubouchi, "Ways of Student Counseling's Approach Lead to Prevention for New Students to Drop out University," *J. Yamanashi Eiwa Coll.*, vol. 8, pp. 9–17, 2009.

[8] S. Fukushima, "A case study of the Comprehensive Student Information Data Analysis System : Enrollment management and institutional research at Yamagata University," *J. Inf. Process. Manag.*, vol. 58, no. 1, pp. 2–11, 2015.

[9] K. Yamamoto, "IR for Effectively Advancing Enrollment Management," *Res. Bull. kyoto koka women's Univ.*, vol. 51, pp. 89–98, 2013.

[10] T. Hashimoto, "Model construction of university students' preliminary dropout prevention measures (IR) based on data: Focusing on differences in institution between Japan and the US," 2015. [Online]. Available: https://kaken.nii.ac.jp/ja/report/KAKENHI-PROJECT-15H00090/15H000902015jisseki/. [Accessed: 15-Mar-2018].

[11] K. Takamatsu, K. Murakami, K. Takao, J. Asahi, T. Kirimura, K. Bannaka, I. Noda, K. Mitsunari, T. Nakamura, and Y. Nakata, "Probability prediction of dropout of students on EMIR," *Proceeding 7th Meet. Japanese Institutional Res. MJIR2017*, pp. 60–65, 2017.

[12] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: A structure for efficient numerical computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, 2011.

[13] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 99–104, 2007.

[14] F. Pedregosa and G. Varoquaux, *Scikit-learn: Machine learning in Python*, vol. 12. 2011.

[15] W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python Sci. Conf.*, vol. 1697900, no. Scipy, pp. 51–56, 2010.

[16] D. Wolpert, "No free lunch theorems for search," *Most*, pp. 1–38, 1995.

[17] D. . Cox, "The Regression Analysis of Binary Sequences," *J. R. Stat. Soc.*, vol. 20, no. 2, pp. 215–242, 1958.

[18] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.