IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 003, LIR160
DOI: https://doi.org/10.52731/lir.v003.160

# To Predict First-year Grades by Including Data on Relationships with People Inside and Outside the University

Naruhiko Shiratori *

## Abstract

In this study, in order to predict first-year credits, an important indicator for preventing students from dropping out of college, we created a model using three types of data: basic data before entering college, relationship data with non-students, and credit data, and conducted comparative verification for each combination of the three types of data. In the experiment, the accuracy of the model improved as the number of data was increased from basic data, relationship data, and unit data. In order to accurately predict students at high risk of dropping out, the number of credits for the spring semester of the first year should be available after the number of credits is available, but in order not to overlook high-risk students, we used a logistic regression model with an awareness of recall and showed that prediction is possible to some extent using basic data and relational data.

*Keywords:* dropout, predicting performance, relation data, random forest, logistic regression

## 1 Introduction

In this study, we will construct a model that incorporates both student data and data on people in contact with students inside and outside the university in order to predict first-year credits earned, which is an important indicator for preventing students from dropping out of the university. Previous studies on dropout prediction have focused mainly on the use of student data, and have rarely included the relationships and influences of parents, friends, and other people inside and outside the university. In Europe and the U.S., studies on the relationship between first-generation research and dropout have been accumulating as data on relationships, but in Japan, limited predictive research is still the main focus. First-generation studies and other studies have shown that data on related parties other than the student in question, such as parents and friends, such as educational background and occupation, are related to the student's success. In this study, the relationship data from both inside and outside the university were obtained from the enrollment questionnaire and used as explanatory variables along with the number of absences and grade point average in high school and the number of credits in the spring semester of the first year as basic data possessed by students before college, to create a predictive model for high-risk students. Another model was created to predict student success by obtaining the number of first-year credits as student success. In validating the model, Accuracy, Recall, Precision, and F1 were examined for accuracy, and random forest and logistic regression were used.

* The Faculty of Management and Economics, Kaetsu University, Tokyo, Japan

## 2 Related Works

### *A. Relationship between First-Year Credit Count and Withdrawal from College*

College dropout is directly linked to managerial challenges for universities, and for students, it leads to problems that waste their time and can easily lead to career problems. Therefore, it is necessary to detect students who are likely to drop out (high-risk students) at an early stage and implement dropout prevention measures at the appropriate time.

One of the key factors in detecting whether a student is a high-risk student is the number of credits and grades earned while in school, and Bonifro uses a combination of three data groups to predict students who will drop out in their first year: basic information such as gender and age group, the need for additional study requirements, and credits, and conducts evaluation. It states that accuracy is better when two sets of data, basic data and learning requirements, are combined than when only basic data are used, and accuracy is even better when forecasts are made using all the data plus the number of units. Although accuracy is better with more data, he states the advantages of being able to make risk determinations for students even at the application stage, and the flexibility of being able to make predictions in a way that can be updated after the first year is completed [1].

Fig.1 shows the number of credits in the spring semester of the first year for students who withdrew and those who did not withdraw at the universities studied. The average number of credits for students who withdrew in the spring semester of their first year was 11 credits, while the average number of credits for the other students was 18 credits; in the spring semester of the first year, the difference was 7 credits, but by the end of the first year, the average number of credits for students who withdrew was 24 credits and the average number of credits for those who did not withdraw was 36 credits, widening the difference to 12 credits. As mentioned above, previous studies have shown that the number of credits and grades are related to withdrawal. Early prediction of the number of credits and grades, and linking them to effective measures, will prevent dropouts.

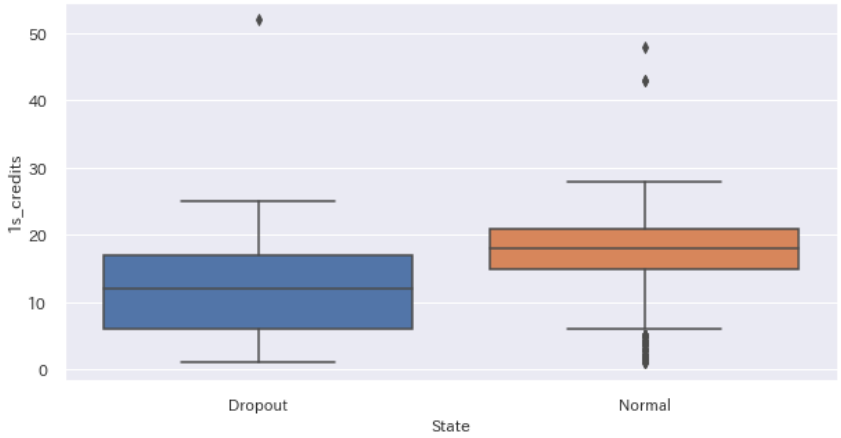


Figure 1: Number of first-year spring semester credits of dropouts and graduates

### B. Relationship between Student Affiliates and Performance

Takeuchi presents a model of student socialization in which parents' social class, educational background, and educational expectations are related to the student's attributes, high school, etc., and then to subsequent college experiences, post-university occupations and lifestyles. In addition, he points out that students' pre-college characteristics are not limited to academic ability and knowledge, but also include study habits, reading, and dating/relationships, and that these characteristics persist even after entering university, pointing out the need to prepare students for the new society of university [2]. Kono reviews other first-generation studies in the U.S. and Japan from NCES, noting that first-generation students at four-year universities have lower GPAs and higher rates of remedial education, but that the gap between them and other students narrows as they prepare to enter college [3]. Furthermore, a survey conducted at a P university in Japan revealed that first-generation students wanted to quit college and had problems with their studies. As described above, previous studies in the U.S. and Japan have shown that first-generation students have some learning characteristics compared to other students.

### C. Predicting First-Year Performance

Many studies of dropout prediction in universities include variables that are defined prior to enrollment. For example, in a review paper written by Hellas et al. in a study predicting student performance [4]. The paper summarizes the three main focuses in forecasting research: what are the variables to be predicted, what are the variables to be entered, and what is the methodology? Various studies are presented in this review paper, and a quarter of the papers that measure student performance use course grades, or scores, as the objective variable. In this paper as well, we will use the number of credits earned as one of the scores to predict high-risk students. Shiratori et al. introduce the number of days absent from high school as a pre-enrollment variable as an explanatory variable for predicting dropout and confirm its significance [5]. Furthermore, they show the number of credits earned in the spring semester of the first year as an important variable for predicting dropout. In this study as well, using the variable that can be obtained prior to enrollment and the number of credits earned in the spring semester of the first year, we used the number of credits earned in the first year.

In Japan, there have been few studies on risk prediction of students connected with human data other than those represented in the first-generation studies. In this study, we will examine the extent to which human factors enter into risk prediction by comparing patterns when variables intended for human relationships are included and when they are not included in the variables of the prediction model.

## 3 Methods

### A. Datasets

The data used in this study is the enrollment data for the academic years 2019 to 2021 at University A. University A is a college of social sciences in the humanities. There were 351 students enrolled in the 2019 academic year, 340 in the 2020 academic year, and 303 in the 2021 academic year, for a total of 904, of which 285 students responded to the new student survey in 2019, 303 in 2020, and 290 in 2021, for a total of 878 (97% response rate), and the experiment used this data were used.

### *B. Variables*

The number of credits acquired in the first year was used as the predictor variable (objective variable) and was divided by the number of credits acquired in the first year (24 credits or less/25 credits or more). 24 credits was used as the criterion for high-risk students because, as mentioned in Chapter 2, the average number of students who dropped out at the universities surveyed was 24 credits. The reason for using 24 credits as the criterion for high-risk students is that the average number of credits earned by students who withdrew from the surveyed universities is 24 credits.

Three types of explanatory variables were used: 1. basic variables known prior to enrollment, 2. variables representing relationships, and 3. credits. A group of variables was used. 2 variables representing relationships were two types of data: the type regarding parents and how they feel about relationships. 2 data was obtained in the enrollment questionnaire in a required class after enrollment. 3 credits were used for the number of credits in the spring semester of the freshman year. Variables in 2, where the variable value was not numeric, were converted as numeric values using One-Hot-Encoding before being used.

Table 1: Relationship Variables

| ***Type*** | ***Feature*** | ***Value*** |
|---|---|---|
| Parent | Q28 Who is the primary household supporter? | Father, Mother, Yourself, Others |
| Parent | Q29 Occupation of the primary supporter | Worker, private business, corporate business, free enterprise, agriculture, forestry, fisheries, other, no occupation |
| Parent | Q30 Household's last level of education | Completed graduate school, graduated from college, graduated from junior college, graduated from vocational school, graduated from high school, other/not sure |
| Relation | Q39 Lively and diplomatic | Completely different, Approximately different, Slightly different, Neither, Slightly agree, Fairly agree, Strongly agree |
| Relation | Q40 Complaints about others, prone to arguments | Completely different, Approximately different, Slightly different, Neither, Slightly agree, Fairly agree, Strongly agree |
| Relation | Q44 Reserved and quiet | Completely different, Approximately different, Slightly different, Neither, Slightly agree, Fairly agree, Strongly agree |
| Relation | Q45 Considerate and kind to others | Completely different, Approximately different, Slightly different, Neither, Slightly agree, Fairly agree, Strongly agree |

### *C. Evaluation Indicators*

The evaluation method used Accuracy, Recall, Precision, and F1 indices for accuracy. Since this is a classification problem, accuracy is often tested using a combination of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Positive here is the prediction that the student is a high-risk student, and Negative is the prediction that the student is not a high-risk student; Negative (FN) Accuracy is calculated as (TP+TN)/(TP+FP+TN+FN), where Recall is TP/(TP+FN ) and Precision is calculated by TP/(TP+FP), while F1 is calculated by the harmonic mean of Recall and Precision. When calculating the precision, a division was made with K=10, 70% was applied to the training data and 30% to the test data for cross-validation, and the average of 10 times was used as the respective precision.

## 4 Experimental Results and Discussion

### *A. Experimental environment*

Two types of models were used, a logistic regression model (LR) and a random forest model (RF), due to multicollinearity considerations and the possibility of interpreting explanatory variables. The experimental environment was Python (version 3.7) and the machine learning environment was Scikit-Learn (version 0.23.2). Variable groups were normalized, and because the objective variable was an unbalanced data set, the experiment was conducted after oversampling using SMOTE [5]. The parameters of the logistic regression model were as follows: penalty was done in L2, C was set to 1.0, and the other default values of Scikit-Learn were used; the parameters of the random forest were as follows: n estimators was set to 100, max depth was not set, and the other default values of Scikit - Learn default values were used.

The following three patterns of experiments were conducted using the following three groups of explanatory variables: 1. basic variables known prior to enrollment, 2. using variables representing relationships, and 3. spring semester units of first-year students.

- 1: Only basic variables known prior to enrollment
- 1+2: Basic variables known prior to enrollment + variables representing relationships
- 1+2+3: Basic variables known prior to admission + variables representing relationships + number of credits for spring semester of freshman year

### *B. Experimental results*

The results of the experiment are summarized in Table 2 below, with Accuracy values summarized in Figure 2, Recall values in Figure 3, and Precision values in Figure 4. Table 2 summarizes Accuracy values, Recall values, Precision values, and F1 values for each experiment and each model.

Table 2: Experimental Results

| Variables | Model | Accuracy | Recall | Precision | F1 |
|---|---|---|---|---|---|
| 1 | RF | 0.757 | 0.281 | 0.129 | 0.176 |
| 1 | LR | 0.583 | 0.502 | 0.129 | 0.205 |
| 1+2 | RF | 0.896 | 0.067 | 0.267 | 0.102 |
| 1+2 | LR | 0.633 | 0.467 | 0.122 | 0.192 |
| 1+2+3 | RF | 0.920 | 0.417 | 0.643 | 0.491 |
| 1+2+3 | LR | 0.903 | 0.783 | 0.493 | 0.601 |

For Accuracy, the accuracy increases as the number of explanatory variables increases from 1, 1+2, to 1+3. Comparing Random Forest (RF) and Logistic Regression (LR), RF is more accurate for all variables in Accuracy, while LR is more accurate for all variables in Recall. In Precision, there is no difference between LR and RF for variable group 1, but RF is more accurate for the other variable groups.

Random Forests generally shows higher accuracy for Accuracy and Precision, while Logistic Regression shows higher accuracy for Recall. One reason for this may be that the logistic regression model, which has linearity with respect to Recall, can select students who tend to be high-risk while still allowing for FN. On the other hand, a random forest, which is a nonlinear model, is better for extracting high-risk students with higher accuracy in overall Accuracy and Precision.

Next, we consider whether the relational data could contribute to the predictive model. This can be considered in terms of the difference between variable group 1 and variable group 1+2. The model with this difference is the Random Forest model of Accuracy and Precision. In order to use more detailed relationship data as variables when determining that a student is a high-risk student, a nonlinear model is more accurate than a linear model. It can be seen that it is better to use relational data when more Precision accuracy is required for prediction.

The three main findings of this experiment are as follows.

- The most accurate way to predict high-risk students is to use the number of first-year spring semester credits.
- To increase the accuracy of Precision and Accuracy, it is better to use human relationship data as a variable and to use a nonlinear model.
- To increase the accuracy of Recall, it is better to use a linear model, but not to use human relationship data.

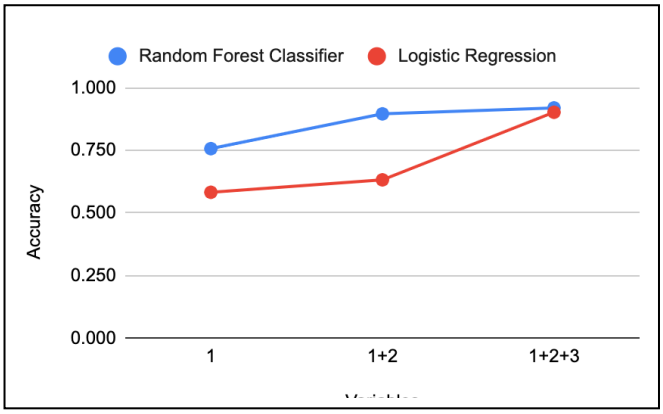


Figure 2: Accuracy Graph

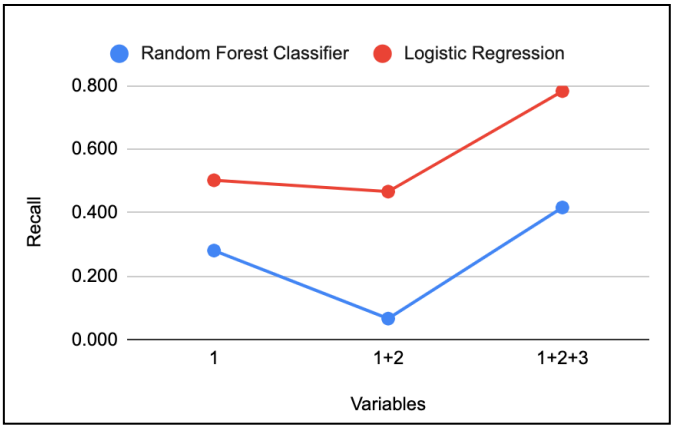


Figure 3: Recall Graph

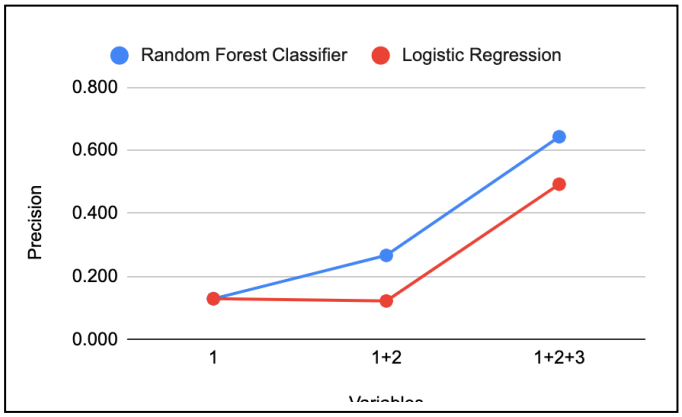


Figure 4: Precision Graph

## 5 Conclusion

In this study, a model was created using basic variables known prior to enrollment, relationship variables, and number of credits to predict the number of credits earned by first-year students, which represents an important factor in college student dropout. While the number of credits is a major factor in improving prediction accuracy, we found that human relations data can also be a factor in improving accuracy by using nonlinear models such as random forests.

On the other hand, it is necessary to ascertain why human relationship data is a factor that increases accuracy by using Feature Importance and other methods to ascertain how the variables contribute to the forecast. We also believe that clarifying how human relationship data relates to the number of credits and grades will enable us to utilize the data in dropout prevention measures.

## Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research C 20K02618.

## References

[1] *F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student Dropout Prediction," in Artificial Intelligence in Education, 2020, pp. 129–140.*

[2] 武内清., "学生文化の実態と大学教育," 高等教育研究, vol. 11, pp. 7–23, 2008.

[3] 河野銀子, "大学大衆化時代における'First-Generation'の位相," 山形大学紀要 教育科学, vol. 13, no. 2, pp. 127–143, Jan. 2003.

[4] A. Hellas et al., "Predicting academic performance: a systematic literature review," in Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, Jul. 2018, pp. 175–199.

[5] 白鳥成彦, 大石哲也, 田尻慎太郎, 森雅生, and 室田真男, "中退確率の遷移を用いた中退学生の類型化," 日本教育工学会論文誌, vol. 44, no. 1, pp. 11–22, 2020.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 200