THE JOURNAL OF KOREAN ASSOCIATION OF COMPUTER EDUCATION Volume 27, Issue 6, 2024

The Journal of Korean Association of Computer Education
Vol. 27, No. 6, 2024
https://doi.org/10.32431/kace.2024.27.6.009

# Predicting Student Loan Defaults in South Korea Using Machine Learning: Insights obtained from SHAP Analysis on KOSAF Data

Doun Jeon[†] · Hansung Kim[††]

†Regular Member Graduate School of Interdisciplinary Information Studies, The Cyber University of Korea
††Lifetime Member Professor of Software Engineering and SW Education, The Cyber University of Korea (corresponding author)

**ABSTRACT**

For students to avail of educational opportunities without discrimination, it is necessary to promote customized policies for vulnerable groups at substantial risk of insolvency, rather than limiting the targets of student loans. This study proactively identifies the risk of student loan defaults and analyzes the key causal factors for it in order to suggest policies that enhance financial stability. To this end, this study utilizes data from the Korea Student Aid Foundation (KOSAF), a national institution, and, using various machine learning models, constructs a model to predict student loan default. The analysis applies the Random Forest, XGBoost, CatBoost, and LightGBM models, thereafter using SHAP analysis to interpret the factors influencing loan defaults. The key results reveal that the CatBoost model demonstrates superior performance, depending on the type of school and loan. Key risk factors for higher default risk included being a student of humanities, social sciences, Art and Physical Education and education; being male; and securing grades below 80. Conversely, the factors that reduced default risk included studying Medical Science, attending metropolitan universities, having grades above 90, and being under 24. Based on these findings, policies for loan screening and delinquency management are suggested to enhance financial stability. This study emphasizes the use of machine learning techniques and explainable AI (XAI) models to improve the accuracy of student loan default predictions and provide valuable policy insights. This study adds to existing literature by identifying the factors for both direct and income-contingent loan default.

**Keywords** Student loan default; machine learning; ensemble; boosting; SHAP; XAI

| | |
|---|---|
| Submission | 2024.08.19. |
| Review Completed | 2024.09.23. |
| Acceptance | 2024.09.24. |
| Published | 2024.09.25. |

* This paper is a summary and enhancement of a portion of the first author's Master's thesis at the Graduate School of Interdisciplinary Information Studies, The Cyber University of Korea.

## 1. Introduction

Loan systems for student support, as well as student loan recipients, loan and repayment conditions differ from country to country. However, the commonality lies in keeping interest rates for student loans low, or allowing repayment of the principal and interest after employment to ease the burden of repayment while studying. As such, the student loan system is a national financial project, and is accompanied by financial burdens, such as issuance of government bonds.Unlike loans from private financial institutions, student loans have less strict credit conditions and are, thus, relatively highly exposed to risks. According to Scott-Clayton[1], as much as 40% of student loans are predicted to lead to insolvency. In such cases, government input and burden on the taxpayer to compensate for it are inevitable[2].

Therefore, institutional supplementation is necessary to prevent loan insolvency and ensure the sustainable operation of the student loan system. However, to provide educational opportunities to a significant number of students, without discrimination, it is necessary to promote customized policies for vulnerable groups at substantial risk of insolvency, rather than limiting the targets of student loans. In particular, delinquency management and credit recovery support should be provided preemptively to prevent intensified insolvency.

In the United States, a relief program called Saving on A Valuable Education (SAVE) has been established to reduce the repayment burden of student loans; in Korea, policy considerations, such as interest exemption during enrollment and deferral of principal and interest repayment, are provided to the vulnerable. Keen identification of vulnerable groups will help promote customized policies and efficient budget utilization[1, 2].

This study contributed to design a model to predict student loan default and proactively identify the key causal risk factors leading to of student loan insolvency and to identify the main factors that cause default. The development of machine-learning algorithm techniques has enabled fast and sophisticated modeling. The terms of agreement for each student loan account and the personal characteristics of the customers holding the accounts vary widely.

A default prediction model should be able to predict the default closest to the actuality when entering various characteristic combinations. Therefore, this study aims to design an excellent prediction model, using various machine learning models and student loan default data.

Even if a model can predict default accurately, it is difficult to interpret the results unless the factors with the maximum influence on prediction are known. In recent credit rating models, an explanatory AI (XAI) model was introduced to explain why credit ratings have changed[3].

Therefore, this study also contributed to ensure the interpretability of the model by revealing the key factors of default after model design.

## 2. Literature review

### 2.1 Direct Lean & Income-Contingent Loan

Since the importance of national fiscal soundness is increasing and education finances also occupy a large sector of national finances, it is worth conducting research on the factors of student loan default. Examining the student support systems of major countries, student loans are divided into two categories, i.e., Direct Loan (DL) and Income-Contingent Loan (ICL)[4].

DL are operated in the United States, Germany, Korea, and Japan, and are similar to loans handled by financial institutions, such as banks, with the repayment period after the loan being set. If the scheduled repayment is not repaid within the deadline, it is classified as overdue. In some cases, debt collection may proceed. ICLs were first introduced in Australia in 1989 and are being operated in the United States, New Zealand, Korea, and Japan. Repayment begins when the borrower earns a certain level of income. Unlike direct loans, these have the advantage of letting the students concentrate on their studies without having to pay off the loans while attending school, when no income is generated.

Even if the borrower becomes insolvent, the repayment of ICL is suspended until graduation, so the insolvency is revealed after a sizeable time. In Korea, ICL borrowers, who have repaid less than 5% of the principal amount five years after graduation, are separately managed as “long-term not repaying borrowers," but there are also limitations limited to graduates. There can be a wide variety of economic changes for students attending or taking a leave of absence, and the risk of insolvency can be responded to early only when these changes are recognized in a timely manner. This study differs from others by identifying the factors for both DL and ICL default.

Predicting and preemptively preparing for poor student loans has developed into an academic field called risk management in financial institutions, including banks. In the event of loan insolvency, financial institutions impose delayed compensation and statistically analyze credit information on individuals to quantify the possibility of credit risk [5]. Thus, loans to low-credit customers are blocked in advance, or a higher interest rate is charged to prevent the expected risk.

### 2.2 Related Studies

Literatures related to the subject of the current study are divided primarily into four categories.

The first relates to factors contributing to student loan defaults. Many studies have shown that demographic factors, such as race, gender, age, and income, have a significant relationship with the level of student loan debt and the difficulty in repaying it[6-12]. Studies have also been conducted on the default rate of student loans by the type of educational institution. For-profit and other schools exhibit higher student loan default rates than traditional public and non-profit private schools. Han & Jeong[13] also developed a model for predicting the number of delinquencies in Korea by applying a logistic regression model according to the period people were delinquent in direct loans. This study identified vulnerable groups in delinquency, such as graduate students and humanities students. Additionally, the researcher mentioned that it is necessary to develop a model that includes additional variables, such as the financial information of delinquents, and to compare its performance with that of other models.

The second is related to 'credit rating using non-financial information. Several studies have proven the effect of using non-financial information on personal credit ratings. Viani B. Djeundje et al. [14] analyzed whether alternative personal information, including gender and age, e-mail sending time, and psychological test results, affect default, using data from Lenddo, the world's first alternative credit rating company. It was found that some variables significantly affect default; in particular, the predictive power (AUC) of the model increased by up to 17%, with the use of the above-mentioned variables, compared to that of individual data. Young-Jun Kwon et al. [15] conducted a study, using additional data on electricity rates, i.e., non-financial information, in their personal credit rating model. Customers who regularly paid electricity bills without delinquency, but did not engage in financial activities, were classified as unevaluable or high-risk groups in the existing credit rating model; therefore, credit loan approval was impossible.

The third constitutes a study of credit rating using machine-learning and deep learning. The development of big data analysis technology has helped researchers enhance the accuracy of default predictions using machine learning. Pedro et al. [16] applied XGBoost to 30 annual financial ratio datasets of 156 commercial banks from 2001 to 2015 to predict bank defaults in the United States. They revealed that lower values for retained earnings to average equity and pretax return on assets increase the risk of bankruptcy. Joo Wan Park et al.[17] used the decision tree, logistic regression, artificial neural network and random forest models, and support vector machine to build a credit rating model for small business owners. By comparing the correct classification rate and F-1 score of each model, using the evaluation data, they found that the logistic regression model had the best prediction performance. Kim & Moon [18] constructed various corporate default prediction models, using corporate data provided by the Taiwan Economic Journal, and compared their classification performance. Apart from individual classification models, the researcher also used ensemble models, such as XGBoost and LightBGM, and found that the classification performance of LightGBM was slightly better.

The fourth is related to the eXplainable AI (XAI). It is necessary to explain the prediction results in order to identify the main default factors. Therefore, researchers are actively developing explicable AI. Moscato et al. [19] constructed a personal credit rating model by applying logistic regression analysis and random forest to lending club data. Subsequently, they demonstrated the effect of the independent variables on the model, using LIME and SHAP. Chun et al.[3] described the factors that significantly influence credit ratings for individuals using Lending Club data and SHAP (Shapley Additive Explanation) values, based on the XGBoost model. In the logistic regression model, the feature importance of the default of each independent variable can be easily identified from the coefficients. However, it is difficult to verify the feature importance of other AI models without the SHAP values. Therefore, it is necessary to identify these factors using XAI to derive future policy implications.

As can be observed, student loan default and AI models have been investigated separately by several researchers, but few studies have combined the two topics.

Therefore, for accurate analysis, it is necessary to apply various machine learning techniques rather than statistical modeling to predict student loan defaults. Additionally, unlike DLs, ICLs are expected to have a higher risk of default due to the uncertainty of future repayments. Based on previous literature, this study attempts to derive detailed and comprehensive implications by considering current student loan issues.

## 3. Method

### 3.1 Data collection

This study aims to develop a prediction model for DLs and ICLs from the Korea Student Aid Foundation (KOSAF), a public institution for student aid in Korea, and to identify the main factors of default. A number of loan accounts, i.e., 6,551,349, were analyzed and the structure of the data is shown in Table 1. All the variables used were categorized as data. Tuition level and student information by school system (college, university, graduate school) and type of school foundation (national, public, and private) differed. Because the data were sufficiently large, they were separated and analyzed for accuracy.

Table 1. Structure of Analysis Data

| No | Feature Name | Elements | No | Feature Name | Elements |
|---|---|---|---|---|---|
| 1 | Loan Classification | 1. Direct Loan | 2 | Type of School Foundation | 1. National and Public |
| | | 2. ICL | | | 2. Private |
| 3 | School Location | 1. Metropolitan area | 4 | School System | 1. Univ 3-yr or shorter |
| | | 2. Gyeongsang | | | 2. Univ 4-yr or longer |
| | | 3. Jeolla-Jeju | | | 3. Graduate School |
| | | 4. Gangwon | | | – |
| | | 5. Chungcheng | | | – |
| 5 | Academic Field | 1. Humanities & Social science & Education | 6 | School Year | 1. Entrants |
| | | 2. Engineering & Natural Sciences | | | 2. Enrolled students |
| | | 3. Medical Science & Pharmacy | | | – |
| | | 4. Art & Physical Education | | | – |
| 7 | Grades for previous semester | 1. Over 90 points | 8 | Income Sections | 1. Basic livelihood recipient |
| | | 2. Under 90 points | | | 2. Under section 4 |
| | | 3. Under 80 points | | | 3. Under section 8 |
| | | | | | 4. Over section 8 |
| 9 | Marriage Status | 1. Married | 10 | Day and Night | 1. Day |
| | | 2. Single | | | 2. Night |
| 11 | Age | 1. Under 24 | 12 | Gender | 1. Male |
| | | 2. 25~29 | | | 2. Female |
| | | 3. 30~34 | | | – |
| | | 4. 35 or older | | | – |
| 13 | Fund Usage | 1. Tuition fee | 14 | Employment Status | 1. 1 (Y) |
| | | 2. Living expenses | | | 2. 0 (N) |
| | | 3. Conversion | | | – |
| 15 | Default Status | 1. 1 (Y) | – | – | – |
| | | 2. 0 (N) | – | – | – |

The meaning of each variable is as follows: The variable, 'Type of School Foundation,' shows whether the customer's affiliated university was a national or private university at the time of application for student loans. Likewise, the variable, ‘School Location,’ describes the categorized area where the customer’s affiliated university is located at the time of loan application. The ‘School System’ is categorized as follows: junior colleges were classified into three undergraduate years or shorter and universities were classified into four-year undergraduate years or more. The ‘Academic Field’ and ‘Day and Night’ classification code means the division of the customer’s affiliated department and day-night classification at the time of application for student loans. The variable, ‘School Year,’ is categorized into Entrants for first year students, and Enrolled Students for others at the time of application. The ‘Grades for previous semester’ variable represents the average grade of students in the semester preceding the time of application. ‘Income Section’ indicates the income level the customer’s household belongs to at the time of application. The higher its value, the more the converted income. ‘Age’ represents the age of customers at the time of application. ‘Fund usage’ is a variable that distinguishes the purpose for which the student loan will be used, primarily consisting of tuition and living expenses. Among the ‘Fund usage’ values, conversion loans are loans that convert Direct Loans, which were executed at relatively high interest rates in the past, to low interest rates of 2.9%,. ‘Employment status’ indicates whether the customer has commenced mandatory repayment by virtue of receiving an ICL repayment standard income or higher as of the end of 2022; this is used to analyze ICL.

All information, except employment and default, is available at the time of application for loans for each account. Student loan accounts are created for each semester according to the purpose of the loan. Therefore, even for the same customer, detailed information for each student loan account varies over time. If a change in customer status affects default, e.g., a long-term delay in student loans, it will be recognized in the customer’s recent account. However, if a customer’s financial condition deteriorates, accounts executed during better financial condition may also be overdue.

### 3.2 Data Analysis & Instrument

First, data pre-processing was performed. Missing values were identified in employment status data. The proportion of missing values in the total data was very small, at 0.08% for Direct Loans and 0.00% for ICL; therefore, missing values were excluded from the analysis.

Second, oversampling and undersampling were performed because the default value, i.e., the dependent variable, was concentrated in N. When the dependent variable is concentrated in a specific category, the model can be over-fitted for the major category and the predictive power of the model for new data can deteriorate. To solve this unbalanced data problem, this study used both oversampling and undersampling methods. Undersampling is a method that decreases the number of samples in the minority category, which accelerates the learning time. However, this causes a loss of sample data. Oversampling solves the problem of information loss by expanding the number of samples in multiple categories by the number of samples in multiple categories. The model performance was better than that of the undersampling method; however, it can cause overfitting and requires a long learning time[20, 21]. In this study, in consideration of efficiency, such as model learning time, DL with Y default and ICL data were oversampled, respectively, to 50% and 40% of the data in N. Additionally, the default N data were undersampled and adjusted so that the proportions of the two categories were the same.

Third, the study separates the data by loan product, school system, and school foundation. Loan products are divided into DL and ICL, the school system is divided into university - three years or shorter, university - four years or longer, and graduate school, and the type of school foundation is divided into national, public, and private schools. However, the data of national and public universities with three years or less are relatively small compared with those of private schools, which can reduce the significance of the analysis. Therefore, data with three years or shorter were integrated regardless of the school foundation. The ICL for graduates was first implemented in 2022; therefore, limited data were available, but this study included them for consistent analysis. The final data used in this study are shown in gray in Table 2.

**Table 2.** Final data
(this study used grey colored, unit: number of accounts)

| Direct Loan | Univ 3-yr or shorter | Univ 4-yr or longer | Graduates | All |
|---|---|---|---|---|
| National/Public | 237 | 11,626 | 16,068 | 27,931 |
| Private | 20,264 | 78,000 | 74,739 | 173,003 |
| All | 20,501 | 89,626 | 90,807 | 200,934 |

| ICL | Univ 3-yr or shorter | Univ 4-yr or longer | Graduates | All |
|---|---|---|---|---|
| National/Public | 1,133 | 37,191 | 150 | 38,474 |
| Private | 79,324 | 218,439 | 359 | 298,122 |
| All | 80,457 | 255,630 | 509 | 336,596 |

Fourth, one-hot encoding was performed for categorical data analysis. For the measurement of model performance, 30% of the data were separated into test data. The models used were Random Forest, XGBoost, LightGBM, and CatBoost, as described above. The parameters were adjusted using a random search method.

Fifth, the predictive power of the constructed model was measured with the F-1 Score, which is a harmonized average of sensitivity and precision and is a balanced evaluation index that considers model overfitting. Recall that the ratio of accounts predicted by default among actual defaulted accounts can be used for the conservative prediction of default accounts; however, due to the nature of student loans, it was decided to use the F-1 Score instead of sensitivity to prevent excessive collection and promote customized financial support[23]. The calculation method for the F-1 Score is presented in Table 3. and Figure 1.

**Table 3.** Confusion Matrix

| | | Real Category | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted Category | Positive(1) | TP<br>(True Positive) | FP<br>(False Positive) |
| | Negative (0) | FN<br>(False Negative) | TN<br>(True Negative) |

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F-1 = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2\times Precision\times Recall}{Precision+Recall}$$

**Figure 1.** F-1 Score formula

Sixth, SHAP was applied to the model with the highest F-1 Score among the ensemble models used. Subsequently, for each data point, separated by loan product type, school system, and type of school foundation, the size and direction of the importance of the variables were verified using SHAP's summary plot.

## 4. Results

### 4.1. Student loan default rate prediction model

In order to construct a model for predicting the student loan default rate, four machine learning models were developed, based on the loan product type and school system. The machine learning models have the advantages that  The predicted value was calculated using the test data and the F-1 Score was calculated by comparison with the actual value. Although the performance of the models was not significantly different, the Random Forest and CatBoost models for DL, and the XGBoost and CatBoost models for ICL, were relatively superior to the others. The CatBoost model, with an advantage in categorical data analysis, demonstrated a stable performance for all loan products and school system data. Table 4 shows a detailed comparison of the model performance by establishing classification and interdisciplinary systems.

Table 4. The modeling results

| Data | Model | F-1 Score | Data | Model | F-1 Score |
|---|---|---|---|---|---|
| Direct Loan<br>Univ 3-yr or shorter | XGBoost | 78.97% | ICL<br>Univ 3-yr or shorter | XGBoost | 76.34% |
| | Random Forest | 79.11% | | Random Forest | 76.49% |
| | LightGBM | 78.75% | | LightGBM | 76.30% |
| | CatBoost | 79.06% | | CatBoost | 76.35% |
| Direct Loan<br>National and Public<br>Univ 4-yr or longer | XGBoost | 75.44% | ICL<br>National and Public<br>Univ 4-yr or longer | XGBoost | 67.37% |
| | Random Forest | 75.71% | | Random Forest | 67.24% |
| | LightGBM | 75.12% | | LightGBM | 67.12% |
| | CatBoost | 75.66% | | CatBoost | 67.11% |
| Direct Loan<br>Private<br>Univ 4-yr or longer | XGBoost | 73.51% | ICL<br>Private<br>Univ 4-yr or longer | XGBoost | 62.40% |
| | Random Forest | 73.57% | | Random Forest | 62.35% |
| | LightGBM | 73.25% | | LightGBM | 62.34% |
| | CatBoost | 73.60% | | CatBoost | 62.46% |
| Direct Loan<br>National and Public<br>Graduate School | XGBoost | 56.60% | ICL<br>National and Public<br>Graduate School | XGBoost | 90.91% |
| | Random Forest | 57.11% | | Random Forest | 83.33% |
| | LightGBM | 56.90% | | LightGBM | 83.33% |
| | CatBoost | 57.06% | | CatBoost | 83.33% |
| Direct Loan<br>Private<br>Graduate School | XGBoost | 61.07% | ICL<br>Private<br>Graduate School | XGBoost | 77.12% |
| | Random Forest | 61.10% | | Random Forest | 78.15% |
| | LightGBM | 61.06% | | LightGBM | 78.15% |
| | CatBoost | 61.37% | | CatBoost | 80.27% |

The Random Forest model demonstrated the best performance for universities with three or fewer years. The F-1 Score reached a maximum of 79.11, indicating a relatively excellent predictive power. In the case of national and public universities, the Random Forest model demonstrated superior performance for general loans, and the XGBoost model, for ICL. The model performance for DL was 75.71%, which was higher than that for ICL. The CatBoost model performed the best in the case of private universities, with a F-1 Score of 73.60%, and the general loan model was superior to the ICL model. In the case of national and public graduate schools, the Random Forest model was found to be excellent for general loans, while the XGBoost model for ICL was 90.91%, which was higher than that of general loans. However, it was judged that attention to model interpretation was necessary because the ICL for graduate students in Korea was implemented in 2022, and the data were not sufficient. As a result of modeling private graduate schools, the performance of the CatBoost model was the best, and that of the ICL model was 80.27%, which is superior to that of general loans.

## 4.2 Key explanatory factors analysis based on eXplainable AI

To analyze the main explanatory factors of the default rate predicted by the model, this study applied SHAP to the CatBoost model, which showed stable performance for each loan product and a summary plot. Because the size of the dataset was large, 10% of the training data were extracted for plotting. The summary plot was expressed as a point of how each variable affected the SHAP value of the individual predicted values, and the higher the frequency, the thicker the plot. Additionally, they were arranged in descending order of SHAP value. This allows us to determine whether the model generally applies the variable as a factor that increases or lowers the risk of default[21].

Figure. 2 shows the SHAP plot derivation results for the loan products, school systems, and types of school foundations. The summary plot is interpreted on the basis of color and direction. For variable values of Y(1) and Y(0), the colors are red and blue, respectively. The plot distributed in the left (right) direction indicates that the variable lowers (increases) the default rate. For example, for DL borrowers in university with 3 or lesser years, the default rate tends to decrease if the university is in a metropolitan area. If the grade at the time of application is less than 80 points (B), the default rate tends to increase (The full figures are presented in the appendix.).

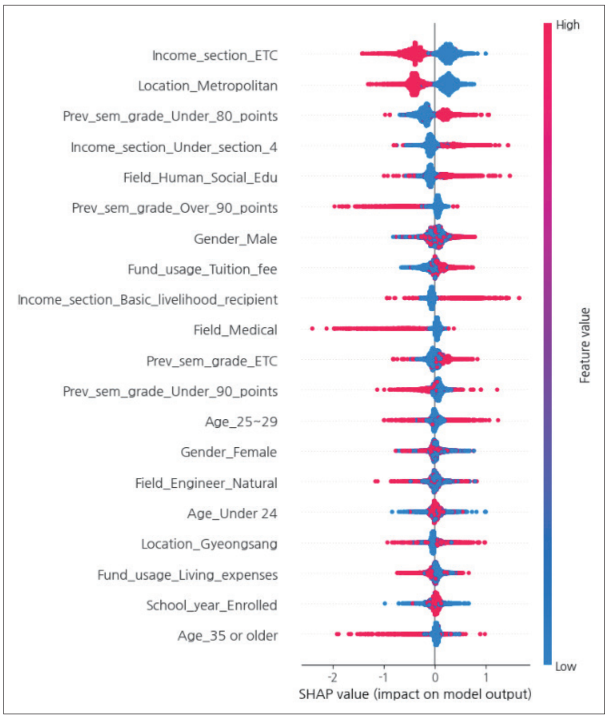


**Figure 2.** Result of SHAP Summary Plot(DL / Univ. 3-yr or shorter)

Figure 3. summarizes the common factors with the maximum effect on the default rate. It was found that for students, under the age of 24, in the medical field, in metropolitan area schools, with grades of 90 points or more, the default rate tends to decline. On the other hand, for students in the humanities and social education/ arts and sports field, male, with grades of less than 80 points, and income deciles of less than four sections, the default rate tends to increase.

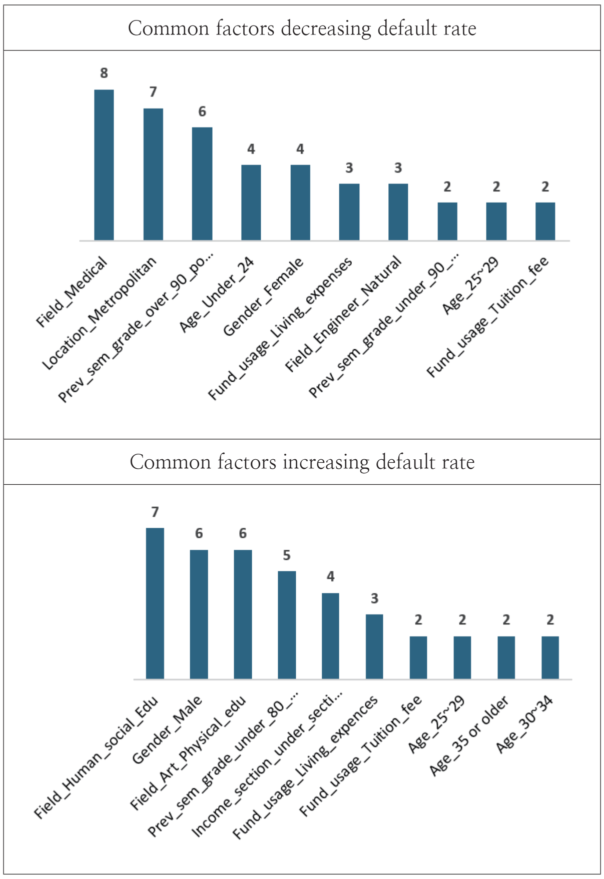


**Figure 3.** Common Factors for Falling and Rising Defaults

Table 5. presents the results of the analysis of the main factors affecting default by loan product and school system type. In the case of universities with a tenure of at most three years, school location and grades appear to significantly influence the default rate, regardless of the loan product. The key factor reducing the default rate is the school being located in a metropolitan area, and that increasing the default rate is the grades in the previous semester being lower than 80 points.

In the case of a DL borrower at a national and public university, the default rate decreases when the grades of the previous semester are 90 points or higher, and increases when the income section is under four. For ICL borrowers, like in case of DL, the default rate increases when the income section is below four. If they receive living expense loans, their default rates decrease.

In the case of a private university student and a DL borrower, the default rate decreases when the grade of the previous semester is at least 90 points, and increases when the score is less than 80 points, similar to the case in national and public universities. In the case of an ICL borrower, the default rate decreases when the age of the applicant is less than 24 years old, and the default rate increases when the income section is less than four or when the major is in the arts and sports field.

In the case of national and public graduate students and DL borrowers, the default rate decreases when the major is in the medical field and the default rate increases when the applicant is over 35 years old. In the case of ICL, the default rate decreases when the age range is 25–29 years. If students are in graduate schools located in the Gyeongsang region or are 30–34 years old, the default rate increases.

In the case of private graduate school students and DL borrowers, the default rate decreases when the major is in the medical field. If the applicant's age is 35 years or older, the default rate increases. For ICL borrowers, if the applicant's age is 30-34 years, the default rate appears to increase.

## 5. Discussion

This study constructs a default prediction model, using student loan default data, and selects the best model to predict the default rate by comparing model performances. Although there were a few differences in model performance by school system and loan product type, overall, the CatBoost model's performance was excellent and stable. Through SHAP analysis, this study reveals the factors with the greatest influence on the default prediction of the CatBoost model and visualizes them through a Summary Plot. The academic implications of this analysis are as follows.

First, in most school systems, the risk of loan default in the humanities, social education, arts, and sports fields is higher than that in other fields. This can be attributed to the income of graduates. According to the 2022 Employment Statistical Yearbook of Graduates of Higher Education Institutions, published by the Korea Education Development Institute, the income of graduates in these fields is lower than that of other departments, such as medicine and engineering. In particular, the arts and sports field, wherein tuition fees are high, is also believed to have increased the risk of student loan default[23].

Second, the risk of default in local colleges was significantly higher than that in metropolitan areas. Students tend to demonstrate high default rates when attending universities in non-metropolitan areas. However, specific school locations, such as the 'Gyeongsang area,' within the non-metropolitan area did not significantly affect default. The results of this analysis indicate the requirement for appropriate policy considerations as the economic capacity of students at local junior colleges is relatively poor and the possibility of insolvency is high.

Third, in the case of universities, grades at the time of applying for a student loan significantly influence  default. Specifically, a score of less than 80 points (grade B) was found to significantly increase the default rate. Each country has different academic requirements for student loans. In Korea, students need 70 points (grade C) or more to use DLs, but ICLs have no restrictions.

The United States also examines whether to continue support based on 2.0 cumulative GPA (grade C) for existing student borrowers(Satisfactory Academic Progress). Therefore, if grade standards are low or not present, it seems necessary to prepare for the possibility of student loan insolvency and the resulting loan losses[24].

Fourth, an age of 24 years or lower for undergraduate, and 25-29 years for graduate students, is a factor that lowers the default rate. People commonly believe that if students borrow money at a younger age, it increases the risk of insolvency by withholding various options for designing the future and incurring additional debt to solve economic problems. Previous studies have shown that the younger the age, the weaker the financial capacity and the higher the risk of insolvency. Other studies that constructed a model for predicting the probability of borrowers' default also found that the younger the age, the higher the probability of default[25, 26]. However, in this study, it was found that being under 24 years old at the time of applying for a loan helps lower the possibility of default. Considering this, it seems that the risk of insolvency is greater if you receive a loan after your usual school age than at a young age. It is presumed that if someone uses student loan after their usual school age, there would be difficulties in continuing studies. For example, he

Table 5. Top 3 Factors of Default Rate by Loan Product and School System

| Direct Loan / Univ 3-yr or shorter | | ICL / Univ 3-yr or shorter | |
|---|---|---|---|
| Decreases default rates | Increases default rates | Decreases default rates | Increases default rates |
| Location_Metropolitan | Prev_sem_grade_under_80_points | Location_Metropolitan | Prev_sem_grade_under_80_points |
| Prev_sem_grade_over_90_points | Income_section_under_section_4 | Prev_sem_grade_over_90_points | Field_Human_social_Edu |
| Field_Medical | Field_Human_social_Edu | Fund_usage_Living_expenses | Fund_usage_Tuition_fee |
| Direct Loan / National and Public / Univ 4-yr or longer | | ICL / National and Public / Univ 4-yr or longer | |
| Decreases default rates | Increases default rates | Decreases default rates | Increases default rates |
| Prev_sem_grade_over_90_points | Income_section_under_section_4 | Fund_usage_Living_expenses | Income_section_under_section_4 |
| Field_Medical | Prev_sem_grade_under_80_points | Age_Under_24 | Field_Art_Physical_edu |
| Location_Metropolitan | Age_25~29 | Prev_sem_grade_over_90_points | Prev_sem_grade_under_80_points |
| Direct Loan / Private / Univ 4-yr or longer | | ICL / Private / Univ 4-yr or longer | |
| Decreases default rates | Increases default rates | Decreases default rates | Increases default rates |
| Prev_sem_grade_over_90_points | Prev_sem_grade_under_80_points | Age_Under_24 | Income_section_under_section_4 |
| Field_Medical | Field_Human_social_Edu | Fund_usage_Living_expenses | Field_Art_Physical_edu |
| Location_Metropolitan | Age_25~29 | Prev_sem_grade_over_90_points | Field_Human_social_Edu |
| Direct Loan / National and Public / Graduate School | | ICL / National and Public / Graduate School | |
| Decreases default rates | Increases default rates | Decreases default rates | Increases default rates |
| Field_Medical | Age_35 or older | Age_25~29 | Location_Gyeongsang |
| Location_Metropolitan | Field_Human_social_Edu | Fund_usage_Tuition_fee | Age_30~34 |
| Gender_Female | Field_Art_Physical_edu | Gender_Female | Fund_usage_Living_expences |
| Direct Loan / Private Graduate School | | ICL / Private Graduate School | |
| Decreases default rates | Increases default rates | Decreases default rates | Increases default rates |
| Field_Medical | Age_35 or older | Age_25~29 | Age_30~34 |
| Field_Engineer_Natural | Fund_usage_Living_expences | Gender_Female | Field_Human_social_Edu |
| Fund_usage_Tuition_fee | Field_Art_Physical_edu | - | Gender_Male |

or she should have to work and stop their study to earn their living expenses.

## 6. Conclusion

Based on the above discussion, the policy implications necessary to prevent student loan default in Korea are as follows:

First, it is necessary to consider the risk of loan defaults in the fields of humanities, social education, arts and sports, and in local colleges. Institutional considerations include strengthening the deferral of repayment, promoting incentive support, or providing a wide range of job preparation opportunities for students or graduates. Particularly, in the case of local colleges, institutional support seems necessary so that students can conduct stable economic activities in the region after graduation.

Second, it is necessary to apply, or strengthen, grade standards when examining the implementation of student loans. The role of education authorities is not only to support student loans fairly, but also to efficiently operate the student support system for those who are willing to study. In Korea, according to public disclosure of universities, as of 2022, the average grade of four-year college graduates is 91.04 points, and the average percentage of graduates with 80 points or more is 93.49%. Only 6.51% of borrowers have a GPA of less than 80, and this is a major factor in student loan defaults. Student loans with relatively low interest rates tend to strengthen the financial welfare function because of rising market interest rates. To strengthen national financial stability and support those who are willing to study, it seems necessary to raise the minimum grade standard to 80 points for DLs and introduce new minimum grade standards for ICLs.

Third, it is necessary to strengthen loan screening or continuously manage delinquency monitoring certificates for students aged over 24 years at the time of application. The usual academic age established for discontinuing education is about 24 years, and applying for a loan beyond this age inevitably means that there is a history of stopping studying. Further research is needed on the reasons for the suspension; however, if it is due to economic difficulties, it can be understood that the difficult situation remains unresolved and leads to insolvency, such as the delinquency of student loans. Efforts, such as limiting the period of deferment, loan periods, or continuing to demand financial education are expected to be needed to prevent potential insolvency from being realized, leading to a burden on the national treasury.

However, this study has some limitations that need to be addressed in subsequent studies.

First, because it used categorized data, it was difficult to predict defaults accurately as individual data changed within the category. As can be seen from the F-1 score of the models, the highest score for each model was in the late 70% range, and when it was low, it was calculated to be mid-50%. In the future, the quality of the research can be improved by using additional specific data for each account.

Second, Koreans recognize that graduating from college is essential for employment, while that may not be true for other countries. There are cases wherein students attend college even though they do not have the money or willingness to study, resulting in student loan default. Therefore, it is difficult to extend the results of this study to other countries. However, in the case of countries that operate a system similar to that of Korea or have an operational plan, this study can provide valuable data.

## References

[1] Scott-Clayton, Judith. (2018). The looming student loan default crisis is worse than we thought; Economic Studies at Brookings, Evidence Speaks Reports 2 (34). https://api.semanticscholar.org/CorpusID:198757392.

[2] J. Cornaggia, K. Cornaggia, H. Xia. (2018). College student behavior and student loan default. SSRN. http://dx.doi.org/10.2139/ssrn.3287952.

[3] Ye Eun Chun, Se Bin Kim, Ja Yun Lee, Ji Hwan Woo. (2021). Study on credit rating model using explainable AI. *Journal of the Korean Data & Information Science Society, vol 32*, 283-295. http://doi.org/10.7465/jkdi.2021.32.2.283.

[4] June suh Yi. (2021). A Study on Government Guaranteed Loan System and Implications in OECD Countries. *The Comparative Economic Review, vol 28*(2), 127–172.

[5] Che Won Song, Hong Soo Kim. (2021). Fintech Score: The Effect of Fintech Service Data on Personal Credit Assessment. *Journal of Information Technology and Architecture, vol 18*(3), 239-253. http://doi.org/10.22865/jita.2021.18.3.239.

[6] Choy, S. P., and Li, X. (2006). Dealing with debt: 1992-93 bachelor's degree recipients 10 years later. *National Center for Education Statistics, vol 2006-156*.

[7] Lochner, L. J., Monge-Naranjo, A. (2011). The nature of credit constraints and human capital. *The American Economic Review, 101*(6), 2487-2529. https://doi.org/10.3386/w13912.

[8] Shapiro, Robert J. (2014). The flawed reasoning and evidence for the Department of Education's Gainful Employment Regulation of private, for-profit colleges and universities. University of Georgetown working paper. https://doi.org/10.2139/ssrn.2542574.

[9] Fox, Jonathan J., Bartholomae, Suzanne, Letkiewicz, Jodi, and Montalto, Catherine. (2017). College student debt and anticipated repayment difficulty. *Journal of Student Financial Aid*. https://doi.org/10.55504/0884-9153.1576.

[10] Adam Looney, Constantine Yannelis. (2015). A crisis in student loans? How changes in the characteristics of borrowers and in the institutions they attended contributed to rising loan defaults. *Brookings Papers on Economic Activity (Fall 2015)*, 1-68. https://doi.org/10.1353/eca.2015.0003.

[11] Adam Looney, Constantine Yannelis. (2018). The consequences of student loan credit expansions: Evidence from three decades of default cycles. University of Chicago working paper. https://doi.org/10.21799/frbp.wp.2019.32.

[12] Luis Armona, Rajashri Chakrabarti, Michael F. Lovenheim. (2018). How does for-profit college attendance affect student loans, defaults, and earnings?. Staff Report, No. 811. https://doi.org/10.3386/w25042.

[13] Jun-Tae Han, Jina Jeong. (2016). Developing the credit risk model for overdue student direct loan. *Journal of the Korean Data & Information Science Society, vol 27*(5), 1293–1305. https://dx.doi.org/10.7465/jkdi.2016.27.5.1293.

[14] Viani B. Djeundje a, Jonathan Crook a, Raffaella Calabrese a, Mona Hamid. (2020). Enhancing Credit Scoring with Alternative Data. *Expert Systems with Applications, vol 163*. https://doi.org/10.1016/j.eswa.2020.113766.

[15] Young-Jun Kwon, Jaihyun Nahm, Min-Jeong Cho. (2011). Economic Effects of Non-Financial Data Sharing.

*Hanguk-gyeongjae-yeongu, vol 29*, 81-107.

[16] C. Pedro, F. Climent, A. Momparler. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economic & Finance, vol Volume 61*, 304-323. https://doi.org/10.1016/j.iref.2018.03.008.

[17] Joo Wan Park, Jin Sung Bae, Hyuk Jun Yoon. (2019). A Study on the Establishment of a Credit Rating Model for Small Businesses Using Big Data Analysis Techniques. KOREG Research report 2019-02.

[18] Jihong Kim, Nammee Moon. (2022). A Study on Classification Models for Predicting Bankruptcy using XAI. *Korea Information Processing Society*, vol 29.

[19] Vincenzo Moscato, Antonio Picariello, Giancarlo Sperlí. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, vol 165*. https://doi.org/10.1016/j.eswa.2020.113986.

[20] Wooseob Yun, Myoung Jong Kim. (2021). AUROC-based Ensemble Model for Bankruptcy Prediction. *Journal of SME Finance, vol41*(3), 41-60. http://doi.org/10.33219/jsmef.2021.41.3.002.

[21] S. M. Lundberg, G. G. Erion, S. I. Lee. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. arXiv Preprint arXiv:1802.03888. https://doi.org/10.48550/arXiv.1802.03888.

[22] Y. Wang, Y. Zhang, M. Liang, R. Yuan, J. Feng, J. Wu. (2023). National student loans default risk prediction, A heterogeneous ensemble learning approach and SHAP method. *Computers and Education: Artificial Intelligence, vol 5*. https://doi.org/10.1016/j.caeai.2023.100166.

[23] Department of Education, Korean Educational Development Institute. (2022). Statistical yearbook for employment of higher education graduates.

[24] Satisfactory Academic Progress, from https://www.ecfr.gov/current/title-34/subtitle-B/chapter-VI/part-668/subpart-C/section-668.34

[25] Dong Gull Lee, Sung In Jun, Jae Wook Chung, Dong Jun Byun. (2014). A Study of the Delinquency Decision Factors and Vulnerability of the Korean Households with Debts. *Journal of Money & Finance, vol 28*(2), 137-178.

[26] June suh Yi. (2019). Analyses on Loan Behavior of Households and Estimation of Household Potential Default Probability. *The Korean Journal of Finance Management, vol36*(1), 63-94. http://doi.org/10.22510/kjofm.2019.36.1.003.

Doun Jeon

- B.A. in Economics, Korea University, 2014
- M.S. in Informatics, Graduate School of Interdisciplinary Information Studies, The Cyber University of Korea, 2024

⊞ Areas of Interest: Machine Learning, Student Loans, Education Finance, Financial Inclusion

✉ kobluejeon@gmail.com

Hansung Kim

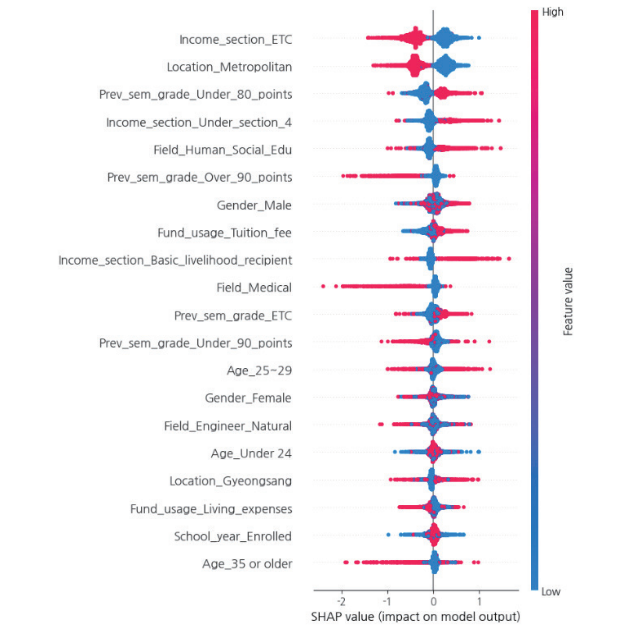
- B.S. in Computer Education, College of Education, Kongju National University, 2005
- Ph.D. in Computer Science Education, Korea University, 2014
- Senior Researcher, Korea Education and Research Information Service, 2013-2020
- Senior Researcher, Software Policy & Research Institute, 2020-2022
- Associate Professor, The Cyber University of Korea, 2022-present

⊞ Areas of Interest: Informatics Education, SW/AI Education Policy, Information Ethics, Digital Literacy

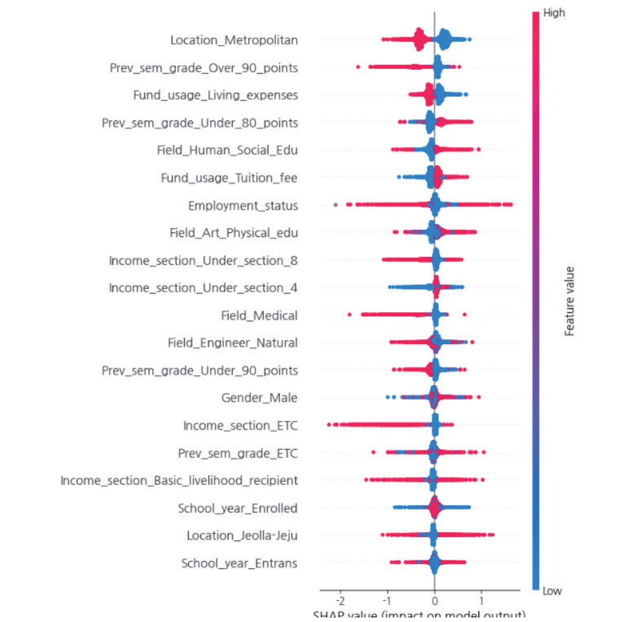
✉ khs4u4u@gmail.com

## Appendix. Full figures of SHAP Summary Plot

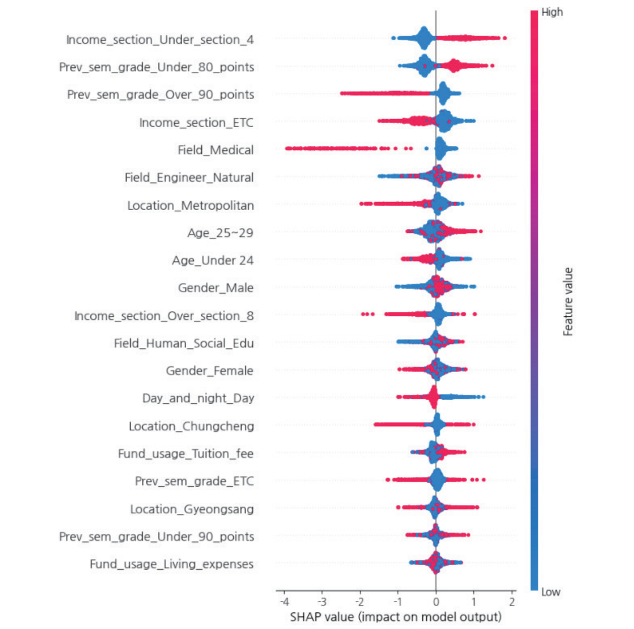
Direct Loan / Univ 3-yr or shorter



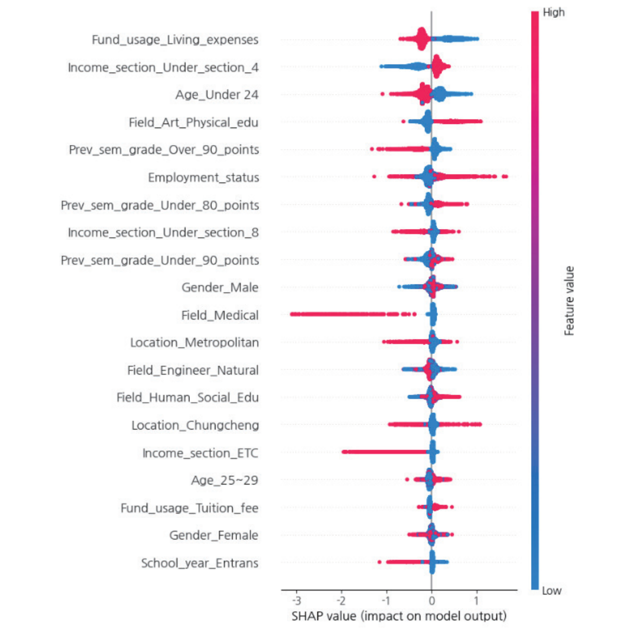
ICL / Univ 3-yr or shorter



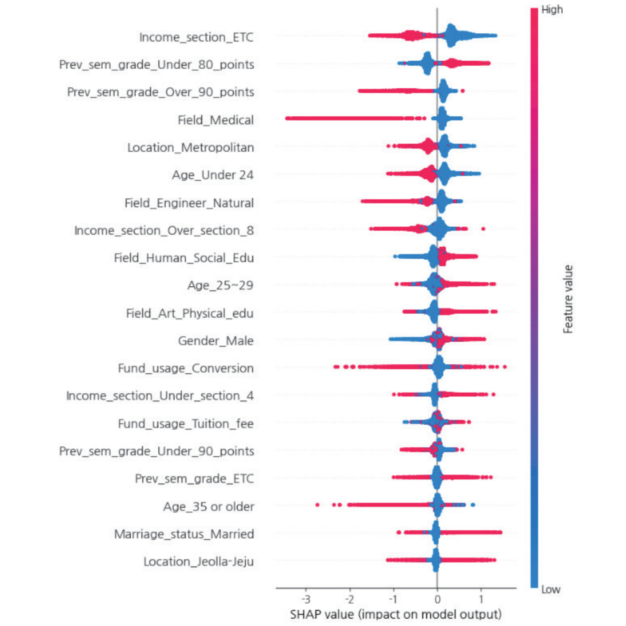
Direct Loan / National and Public / Univ 4-yr or longer



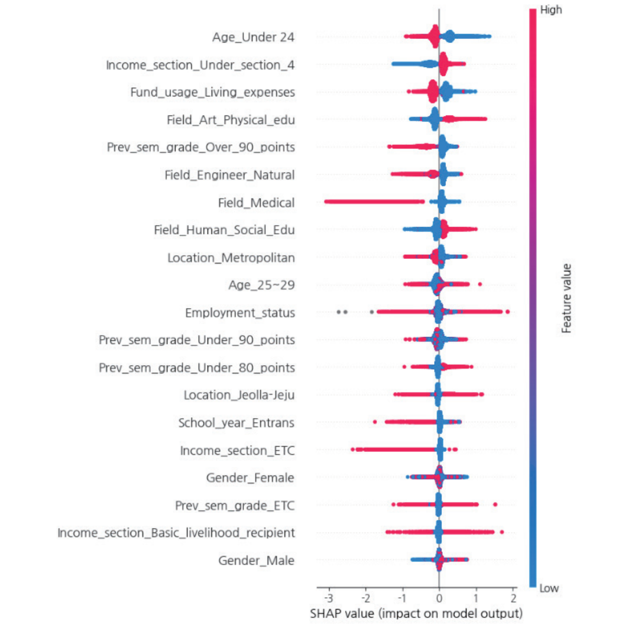
ICL / National and Public / Univ 4-yr or longer



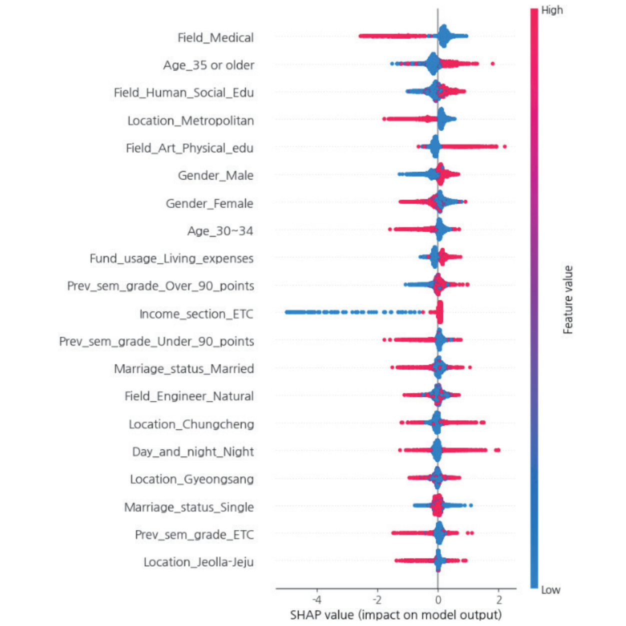
Direct Loan / Private /
Univ 4-yr or longer



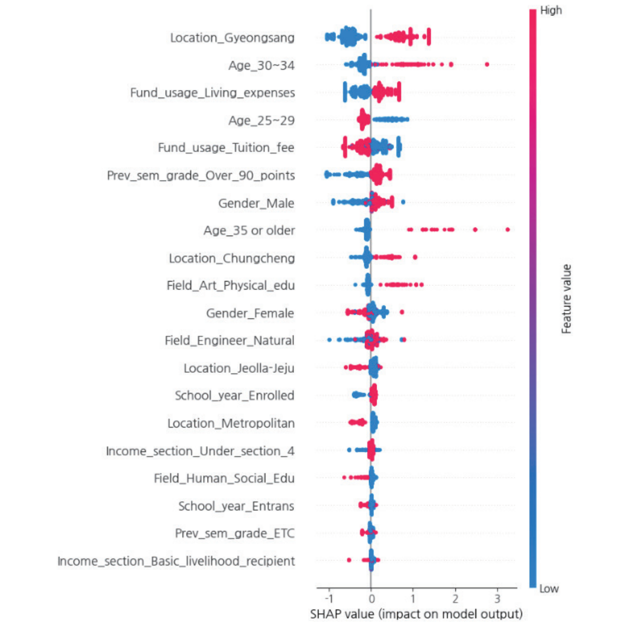
ICL / Private /
Univ 4-yr or longer



Direct Loan / National and Public /
Graduate School



ICL / National and Public /
Graduate School


Location_Gyeongsang
Age_30~34
Fund_usage_Living_expenses
Age_25~29
Fund_usage_Tuition_fee
Prev_sem_grade_Over_90_points
Gender_Male
Age_35 or older
Location_Chungcheng
Field_Art_Physical_edu
Gender_Female
Field_Engineer_Natural
Location_Jeolla-Jeju
School_year_Enrolled
Location_Metropolitan
Income_section_Under_section_4
Field_Human_Social_Edu
School_year_Entrans
Prev_sem_grade_ETC
Income_section_Basic_livelihood_recipient
SHAP value (impact on model output)
High
Feature value
Low