applied sciences

MDPI

Article

The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction

Sunbok Lee [1] and Jae Young Chung [2,*]

[1] Department of Psychology, University of Houston, Houston, TX 77004, USA
[2] Department of Education, Ewha Womans University, Seoul 03760, Korea
* Correspondence: jychung@ewha.ac.kr; Tel.: +82-2-3277-2632

Received: 4 July 2019; Accepted: 29 July 2019; Published: 31 July 2019

Abstract: A dropout early warning system enables schools to preemptively identify students who are at risk of dropping out of school, to promptly react to them, and eventually to help potential dropout students to continue their learning for a better future. However, the inherent class imbalance between dropout and non-dropout students could pose difficulty in building accurate predictive modeling for a dropout early warning system. The present study aimed to improve the performance of a dropout early warning system: (a) by addressing the class imbalance issue using the synthetic minority oversampling techniques (SMOTE) and the ensemble methods in machine learning; and (b) by evaluating the trained classifiers with both receiver operating characteristic (ROC) and precision–recall (PR) curves. To that end, we trained random forest, boosted decision tree, random forest with SMOTE, and boosted decision tree with SMOTE using the big data samples of the 165,715 high school students from the National Education Information System (NEIS) in South Korea. According to our ROC and PR curve analysis, boosted decision tree showed the optimal performance.

Keywords: dropout; machine learning; big data; class-imbalance; oversampling; ensemble

1. Introduction

The negative consequences of students' dropping out of school are significant for both the individual and society. The educational deficiencies of dropout students could severely limit economic and social well-being in their later lives [1]. The society also suffers losses because the nation's productive capacity could be undermined by the shortage of the skilled workforce, and also the dropout students are more likely to be frequent recipients of welfare and unemployment subsidies [2]. Because of those negative consequences, students' dropouts have long been considered as a serious educational problem by educators, researchers, and policymakers. A dropout early warning system can help schools to preemptively identify students who are at risk of dropping out of school and to promptly react to them [3]. The students at risk are likely to drop out without carefully considering the negative consequences of their decisions or without having an opportunity to consult with experts. The early intervention informed by the dropout early warning system can redirect potential dropout students onto the path to graduation and lead them to a better future [4]. Because of the great potential, many governments have developed dropout early warning systems. For example, the department of education and early childhood development in the state of Victoria in Australia developed the Student Mapping Tool (SMT) to help schools to identify students at risk of disengagement and dropout [5], and the state of Wisconsin in the United States developed the Dropout Early Warning System (DEWS) to predict students' dropouts [6]. In the United States, about half of public high schools implemented the dropout early warning systems during 2014–2015 [7].

Machine learning is a promising tool for building a predictive model for a dropout early warning system. However, the class-imbalance could be one of the potential difficulties in implementing a dropout early warning system using machine learning. In the binary outcome representing students' dropouts, the proportions of the two classes (i.e., dropouts and non-dropouts) tend to be imbalanced (e.g., 1.4 percent of dropouts vs. 98.6 percent of non-dropouts in Korean high school in 2016). In general, machine learning classifiers trained on datasets with imbalanced classes tend to show a very poor performance in predicting the minor class because the classifiers ignore the minor class as a noise [8,9]. Because the class of interest in dropout prediction is a minor class (i.e., dropouts), the class-imbalance issue could severely degrade the sensitivity of the early warning system in predicting potential dropout students. Another key issue in the presence of class imbalance is to use performance metrics that are sensitive to performance differences. Traditionally, the area under the curve (AUC) in the receiver operating characteristic (ROC) curve has been widely used in machine learning literature. However, in the presence of class imbalance, the ROC curve analysis may not be sensitive enough to differentiate the performances of classifiers, and the precision–recall (PR) curve may perform better [10]. In the literature, Márquez-Vera et al. [11] predicted student failure at school using genetic programming and different machine learning approaches by addressing the class imbalance issue using the synthetic minority oversampling techniques (SMOTE), and identified the best model based on the true positive (TP) rate, true negative (TN) rate, and accuracy. Knowles [6] used machine learning to build a predictive model of student dropout risk, and identified the best statistical model using the ROC curve. Márquez-Vera et al. [12] used various machine learning algorithms for early dropout prediction, and used the TP rate, TN rate, accuracy, and AUC. However, in the previous literature, the class-imbalance issue and the advantage of using the PR curve have not been fully discussed yet. Thus, the present study aimed to improve the performance of a dropout early warning system: (a) by addressing the class imbalance issue using the SMOTE and ensemble methods in machine learning; and (b) by evaluating the trained classifiers with both ROC and PR curves. To that end, we trained random forest, boosted decision tree, random forest with SMOTE, and boosted decision tree with SMOTE using the big data samples of the 165,715 high school students from the National Education Information System (NEIS) in South Korea. Because the class-imbalance issue is prevalent (e.g., cheaters in online education), the implication of this study is relevant to building predictive models for other educational outcomes as well.

2. Students' Dropouts in South Korea

In this section, we briefly present the current status and reasons for high school dropouts in South Korea to clarify the types of dropouts we are interested in. We also present the rationale for the goal of our dropout early warning system. Table 1 shows the dropout rates of high school students in South Korea from 2010 to 2016 [13]. The dropout rates have decreased from 2.0 percent in 2010 to 1.4 percent in 2016.

Table 1. The Dropout rates of high school students in South Korea from 2010 to 2016.

Year	The Total Number of Students	The Number of Dropouts Students	Dropouts Rates (%)
2016	1,752,457	23,441	1.4
2015	1,788,266	22,554	1.3
2014	1,839,372	25,318	1.4
2013	1,893,303	30,382	1.6
2012	1,920,087	34,934	1.8
2011	1,943,798	37,391	1.9
2010	1,962,356	38,887	2.0

Table 2 shows the reasons for high school dropouts in South Korea in 2013 and 2016 [13]. Students left schools for various reasons: diseases, family problems, poor academic performance, poor relationship with others, strict school rules, and other reasons (e.g., studying overseas and

alternative education). Simply leaving schools does not necessarily mean negative outcomes. Some students leave schools to study overseas, to attend alternative education programs, or to pursue their own career paths earlier. The dropouts due to those positive reasons are not the types of dropouts we are interested in. Chung et al. [14] defined at-risk youths as the youths who are exposed to personal and environmental risks, likely to experience behavioral or psychological problems and find it difficult to achieve normal development without appropriate educational intervention. This group of youths reports high risks of running away from home, dropout, unemployment, violence, prostitution, substance abuse, and other misconducts, crimes, as well as psychological disorders such as depression, anxiety, and suicide. We are interested in predicting the dropouts among those at-risk youths who could benefit from the intervention programs informed by the dropout early warning system.

Table 2. The Reasons for high school dropouts in South Korea in 2013 and 2016.

	Leaving						Expulsion	Total
	Diseases	Family Problem	Poor Academic Performance	Poor Relationship with Others	Strict School Rules	Other Reasons		
2013	1429 (4.7%)	2327 (7.7%)	9887 (32.6%)	486 (1.6%)	1019 (3.4 %)	14,094 (46.5%)	1090 (3.6%)	30,287 (100%)
2016	882 (4.7%)	503 (2.7%)	4047 (21.6%)	222 (1.2%)	225 (1.2%)	11,855 (63.3%)	998 (5.3 %)	18,732 (100%)

The optimal performance of a classifier is only meaningful in relation to a specific task. In a dropout early warning system, the sensitivity represents the proportion of actual dropout students predicted correctly. In this study, we aimed to maximize the sensitivity of our dropout prediction for the following reason. The primary aim of public schools is to support successful learning of all students without a single failure. UNESCO has placed "Education for All (EFA)" as the international policy agenda [15]. The idea behind the agenda is to ensure all students around the world benefit from education. The United States has been building an accountability system to keep all students from falling behind through the "No Child Left Behind Act" and "Every Study Succeeds Act" [16]. To support successful learning of all students, in South Korea, it was proposed to operate a three-tier dropout prevention programs [17]. In the first-tier, all students participate in the general prevention program designed to prevent school dropouts. In the second-tier, students who were identified as being at-risk of dropouts participate in more specialized group- or individual-based prevention programs. In the third-tier, students who express their intention to dropout have opportunities to deliberate their decisions for two weeks before making their final decisions. During the period of deliberation, students receive personalized counseling and training. In this proposed three-tier prevention programs, it is very important to preemptively identify all the potential dropout students and to promptly react to them in order to reduce the number of students who actually want to drop out. Another important performance metric in a dropout early warning system is the precision (or positive predictive value), which represents the proportion of predicted dropout students who actually dropout. The precision is important because it is directly related to the cost for the intervention.

3. Analysis Plan

3.1. Supervised Learning

We used supervised learning in machine learning to train our binary classifiers that predict students' dropouts and non-dropouts. The goal of the supervised learning is to estimate the best mapping function $f(.)$ from the set of features (Xs) to the target label (Y) by training a specific machine learning model on a dataset. The learning process in the supervised learning focuses on assuring the capability of a trained model generalizing knowledge learned from the current observations to

the future observations. The emphasis on the generalizability of the model makes the overfitting a critical issue in supervised learning. The overfitting is said to occur when a model is more complex than necessary and therefore fits too much noise in the training dataset. The model complexity is controlled by the so-called hyper-parameters of the model (e.g., the strength of the penalty in the regularized regression, the depth of a tree in the decision tree). Therefore, the key task in supervised learning is to determine the optimal values of hyper-parameters to make a balance between bias and variance. In practice, k-fold cross-validation is often used to determine (or tune) hyper-parameters. In k-fold cross-validation, the training dataset is partitioned into k equal-sized subsets, each of which is called a fold. At each iteration from 1 to k, a single fold is retained as the validation dataset, and the remaining $k-1$ folds are retained as the training dataset, and the model trained on the training dataset is evaluated on the validation dataset. Then, k performance metrics from the k iterations are averaged to produce a less biased estimate of the performance of the model. The best hyper-parameters can be determined by comparing the averaged performance metrics from k-fold cross-validation of the models with different values of hyper-parameters. We used 10-fold cross-validation for our analysis.

3.2. The Problem of Class Imbalance

The focus of our analysis is to train our binary classifiers by addressing the class-imbalance issue. In classification, a classifier predicts categorical labels (or classes). The classes are imbalanced if there are many more instances of some classes than others in a dataset [18]. The class imbalance is prevalent because many real-world applications, such as fraud detection, spam detection, anomaly detection, and psychological diagnosis, are composed of a large number of normal examples with only a small number of abnormal or interesting examples [19]. The ratio between minor and major classes can be 1:100, 1:1000, or even 1:10,000 depending on the applications. For example, fraudulent cases in retail banking are about 0.1 percent [20].

The class imbalance poses a difficulty in classification because the classifiers trained on the imbalance dataset tend to show a higher predictive accuracy on the major classes, but a poorer predictive accuracy on the minor classes. This bias toward the major classes happens because the performance metrics tend to treat the minor classes as the noise, and therefore guide the classifiers to ignore the misclassification of the minor classes during the learning process [9]. Table 3 shows a hypothetical confusion matrix illustrating the case where the accuracy of a binary classifier can be deceptively excellent when the classifier completely ignores the misclassification of the minor classes. In this example, the proportions of dropouts and non-dropouts are imbalanced, i.e., 100 dropouts (1%) vs. 9900 non-dropouts (99%). The accuracy of the binary classifier in this example is 0.99 despite the complete misclassifications of the dropouts. In this way, when classes are imbalanced, the minor classes are often more misclassified than the major classes. Because the minor class is usually the class of interest in classification, the class imbalance has been a challenging problem in data mining community [21,22].

The problem of class imbalance is a well-known issue in the machine learning community. However, less attention has been paid to this issue when developing dropout early warning system. As presented in Table 1, high school dropout rates in South Korea are less than 2%. According to the National Center for Education Statistics, high school dropout rates in the United States have decreased from 27.2% in 1960 to 6.1% in 2016. Because the class of interest in the dropout prediction is a minor class (i.e., dropout), the issue of class imbalance needs to be properly handled when building a predictive model for the dropout early warning system.

Table 3. The hypothetical confusion matrix showing the case where the accuracy of a binary classifier can be deceptive for an imbalanced dataset. In this example, the proportions of dropouts and non-dropouts are imbalanced, i.e., 100 dropouts (1%) vs. 9900 non-dropouts (99%). Despite of the complete misclassifications of the dropouts, the accuracy is still $0.99 = (0 + 9900)/(0 + 0 + 100 + 9900) = 9900/10000$ because the correct predictions of the non-dropout dominate the accuracy.

		True Labels	
		Dropout	**Non-Dropout**
Predicted lables	Dropout	0	0
	Non-dropout	100	9900

3.3. SMOTE

Many methods have been proposed to address the problem of class imbalance. In the literature, those methods are typically categorized into four groups: algorithm-level, data-level, cost-sensitive, and ensemble approaches [8,9]. The algorithm-level approaches modify existing classification algorithms to be biased toward the minor classes to improve the performance of the model in predicting the minor classes [23,24]. These approaches require knowledge of the specific classifiers, and why the classifiers fail to modify the algorithms. The data-level approaches rebalance the imbalanced classes by over-sampling the minor classes or under-sampling the major classes to alleviate the effect of imbalanced classes [19]. Because these approaches are implemented in the preprocessing step, they are independent of specific classifiers and also can be more easily implemented. The cost-sensitive approaches combine both the data- and algorithm-level approaches by adding high misclassification costs for the minor classes (data level approaches), and also modifying the classification algorithms to accept the costs ([25], algorithm level approaches). Recently, the ensemble approaches are attracting more and more attention as a solution to the class imbalance problem. Hybridizing the bagging and boosting paradigms in ensemble methods with the algorithm-level, data-level, and cost-sensitive approaches for the imbalanced dataset turned out to be very promising. Readers who are interested in a more compressive review on this topic are referred to the works of Galar et al. [9] and Haixiang et al. [8].

In this study, we used SMOTE [19] from the data-level approaches to address the problem of class imbalance. In the data-level approaches, the under-sampling randomly eliminates some major classes to make the major classes less effective on the learning process. However, the under-sampling could remove major classes which are more representative and informative than others, and therefore the decision boundary could be biased. The over-sampling randomly replicates the minor classes to make the minor classes more effective on the learning process. However, the over-sampling creates repeated copies of the same minor class instances many times, and therefore the classifiers could overfit to these minor class instances. After reviewing previous studies on over- and under-sampling, Chawla et al. [19] summarized that the under-sampling showed better performance than the over-sampling, and the combination of the over- and under-sampling did not outperform the under-sampling. They proposed SMOTE as a new over-sampling technique in which the minor class instances are over-sampled by creating synthetic instances rather than creating the same minor instances multiple times with replacement, and showed that the combination of SMOTE and under-sampling performed better than the plain under-sampling. In SMOTE, the minor class instances are over-sampled by creating synthetic instances along the line segments joining the k nearest neighbor instances with minor classes. The idea behind the SMOTE is to over-sample similar instances rather than to over-sample the same instances multiple times so that the classifiers are not overfitted to the minor class instances.

3.4. Ensemble Methods

Machine learning has been successfully applied for predictive modeling in various fields. Especially, ensemble methods have gained considerable attention in recent years because ensemble

methods can substantially improve the accuracy of predictions by combining multiple machine learning algorithms [26]. Each algorithm in the ensemble methods is usually called a base (or individual or component) learner. The ensemble of learners has outperformed a single learner in various tasks, such as the object detection [27], lung cancer identification [28], and fraud detection [29]. The superior performance of ensemble methods mainly comes from the better generalizability of the ensemble of learners.

In ensemble methods, bootstrap aggregating (bagging) and boosting are the two popular paradigms. As the name indicates, bagging improves the accuracy of predictions by aggregating predictions from multiple learners. In bagging, the N base learners are trained in parallel on N bootstrap samples from the original dataset, and then the N predictions from the N base learners are combined using majority voting or other synthesizing methods. Because of its aggregating (or averaging) nature, bagging can reduce the variance of a model. The random forest is a popular machine learning algorithm that uses the bagging paradigm. The random forest is the collection of N decision trees, and combines N predictions from the N decision trees trained on the N bootstrap samples to make a final prediction.

Boosting is another ensemble paradigm. The idea behind boosting is to produce a series of weak learners to make a strong learner. Unlike bagging in which learners are trained in parallel and the predictions are aggregated without preference to any learner, boosting sequentially trains learners, and the predictions are aggregated with heavier weights on better learners. For example, the adaptive boosting (AdaBoost) calls a weak or base learner repeatedly in a series of rounds $t = 1, ..., T$ by weighting previously misclassified instances with higher weight, and then combines the T predictions from the T base learners with the weights of learners determined during the training process [30]. Notice that both the instances and learners have their distributions of weights which are updated across the rounds. The weight distribution for instances at round t makes the learner at round t focus on the instances that were misclassified at round $t - 1$, whereas the weight distribution for learners determines how to combine the T learners after T rounds. In addition to the variance, the boosting can also reduce the bias because of its adaptive nature. In sum, the bagging trains a set of learners in parallel to reduce the variance by exploiting the independence between learners, whereas the boosting sequentially trains a series of learners to reduce both the bias and variance by exploiting the dependence between learners.

In total, we trained four classifiers using the big data samples of the 165,715 high school students from the NEIS database in South Korea: random forest, boosted decision tree, random forest with SMOTE, and boosted decision tree with SMOTE.

3.5. Performance Metrics for Binary Classifiers

A confusion matrix is the cross-tabulation between the true labels from the labeled dataset and the predicted labels from a classifier. Depending on the values of true and predicted labels, the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) can be defined, and these four cases are used to define sensitivity (or recall, TP/(TP + FN)), specificity (TN/(FP + TN)), accuracy ((TP + TN)/(TP + FN + FP + TN)), negative predictive value (TN/(FN + TN)), and positive predictive value (or precision, TP/(TP + FP)).

The area under the curve (AUC) is another popular metric. The AUC is the area under the receiver operating characteristic (ROC) curve [31]. The ROC curve is the curve in the two-dimensional ROC space whose y-axis represents a classifier's true positive rate (or sensitivity) and x-axis represents a classifier's false positive rates (or 1 − specificity). Each point in the ROC space represents a classifier's (false positive rate, true positive rate) pair. If the classifier is the discrete classifier that produces only a class label (e.g., dropout or non-dropout), then the performance of the discrete classifier can be represented as a point in the ROC space. In the ROC space, the upper left point (0, 1) represents a perfect classifier that never predicts the true negative labels as positive (i.e., false positive rate = 0) and predicts all the true positive labels as positive (i.e., true positive rate = 1). Some classifiers

produce the probability of the instances being a member of a class, instead of the discrete class label. Such classifiers can be converted into the discrete classifiers by introducing a threshold. The ROC curve for such classifier is the set of (false positive rate, true positive rate) pairs for all possible thresholds. The ROC curve allows us to visualize the trade-off between the true positive rate (or sensitivity) and false positive rates (or 1 − specificity) of a classifier. That is, it shows that a classifier cannot increase the true positive rate without increasing the false positive rate [9].

Although the ROC curve is a popular evaluation metric for binary classifiers, the precision–recall (PR) curve is recommended when evaluating binary classifiers trained on imbalanced datasets [10]. When the number of negative instances is very large, the true negative (TN) is also likely to be very large. Then, the large true negative (TN) makes the specificity less sensitive. Therefore, the precision (or positive predictive value) could be a more sensitive measure for the imbalanced data because the precision is not affected by the large true negative (TN). Notice that recall is just another name for the sensitivity, and therefore the PR curve only replaces the false positive rate in the ROC curve with the precision. In PR curves, good classifiers aim for the upper right corner. In the present study, we used both the ROC and PR curves for model-wide threshold-free evaluation of binary classifiers.

4. Methods

4.1. Data

Samples. Our sample consists of the big data samples of 165,715 high school students from the NEIS database of 2014. Those are students from two big cities and two provinces in South Korea: Seoul, Incheon, Gyeongsangbuk-do, and Gyeongsangnam-do. The proportion of male students is 0.60. The proportions of freshman, sophomore, and senior in high school are 0.33, 0.34, and 0.33, respectively. The NEIS is the web-based integrated administration system that connects South Korean's education organizations including around 12,000 elementary, middle and high schools, 17 city and provincial offices of education, and the Ministry of Education. The NEIS was developed by South Korea's Ministry of Education in the early 2000s for several purposes. The NEIS was designed to reduce teachers' workload. For example, teachers do not need to prepare various reports, statistics, and administrative documents. In addition, the NIES was designed to enhance the conveniences of citizens, especially parents. For example, parents can request 38 types of student information (e.g., school schedule, meal schedule, grades, and absence) and official certificates online. The NEIS was also designed to enhance the efficiency in school administration. For example, the NEIS reduces manual document preparation, enhances information sharing, and improves decision making processes at a policy level. Currently, the NEIS is maintained by the Korea Education & Research Information Service (KERIS) under the Ministry of Education. The NEIS has two databases. The educational affairs database contains information more than six million students, such as academic achievement, absence, health and so on. The school administration database contains information about HR affairs, teacher information, and school information. Ethics Committee/Institutional Review Board approval for this study was not sought because we used the government data that are already collected in the NEIS system, and IRB approval is not required in South Korea in this case.

Target label. In this study, the target label for prediction is students' dropouts. The binary target label representing students' dropouts was created based on variables named "the school register change" and "the reasons for the school register change" in the NEIS database. The variable named "the school register change" has 15 categories such as entrance, expulsion, leaving, transfer, and graduation. The variable named "the reasons for the school register change"' describes 43 reasons for the changes. As discussed above, we are interested in the dropouts among those at-risk youths who could benefit from the intervention programs informed by the dropout early warning system. Therefore, in this study, we defined dropout students as the ones who have dropped out of school for the 13 negative reasons presented in Table 4. In total, out of 165,715 students, 1348 students (0.81%) were identified as dropout students in our analysis.

Table 4. 13 reasons for the school register changes considered as dropouts.

Category	Specific Reasons	Counts	Percentages (%)
Behavior	Violation of School Rules	126	9.3
Behavior	Requests from Autonomous Committees	15	1.1
Behavior	Assault/Burglary	4	0.3
Behavior	Others	5	0.4
Family	Family Trouble	9	0.7
Maladjustment	Poor Academic Performance	690	51.2
Maladjustment	Victimization	1	0.1
Maladjustment	Relationship with Friends/Teachers	23	1.7
Maladjustment	Strict Rules	66	4.9
Maladjustment	Others	375	27.8
Disease	Disease	8	0.6
Others	Running away from Home	21	1.6
Others	Others	5	0.4
		1348	100%

Features. Following the recommendation from the National High School Center, we used the attendance, behavior, and course performance (the "ABCs") as the key indicators for our dropout predictions [32]. Previous studies also showed that low attendance at the beginning of a semester can be an indicator for the dropout prediction [33]. Therefore, we further included the attendance in the first four weeks as indicators. In sum, we used 15 features to predict students' dropouts: the unauthorized absence in the first four weeks, unauthorized early leave in first four weeks, unauthorized class absence in first four weeks, unauthorized lateness in first four weeks, unauthorized absence, unauthorized early leave, unauthorized class absence, unauthorized lateness, number of self-regulated activities, number of club activities, number of volunteer activities, number of career development activities, normalized ranking on Korean, normalized ranking on Math, and normalized ranking on English.

4.2. Preprocessing, Tuning, Training, and Testing

We used the `caret` package in R to preprocess, tune, train, and test the four classifiers.

Preprocessing. The original dataset that consists of 165,715 students was divided into training (80%; N = 132,573) and testing (20%; N = 33,142) datasets to train and evaluate the four classifiers. For the preprocessing, the 15 features in the training dataset were centered, scaled, and median imputed using the `preProcess()` function in the `caret` package.

SMOTE. The preprocessed training dataset was over- and under-sampled using the `SMOTE()` function in the `DMwR` package in R.

Tuning. The 10-fold cross-validation was used to tune the hyper-parameters of each classifier by setting the method option to `cv` and `number` option to 10 in the `trainControl()` function in the `caret` package. The optimal hyper-parameters were chosen by comparing the classifiers' ROCs with different values of hyper-parameters.

Training. The `train()` function in the `caret` package in R was used to train the random forest, boosted decision tree, random forest with SMOTE, and boosted decision tree with SMOTE.

Testing. The trained classifiers were evaluated on the testing dataset using the `predict()` function in the `caret` package. The testing dataset was not oversampled in any case.

5. Results

Figure 1 presents the density plots for the selected eight features. Each plot presents the density plot of a specific feature for both the dropouts (shaded as red) and the non-dropouts (shaded as blue). Figure 1 shows that the dropout students are more likely to be problematic in attendance and achievement, and are less likely to participate in the school activities.

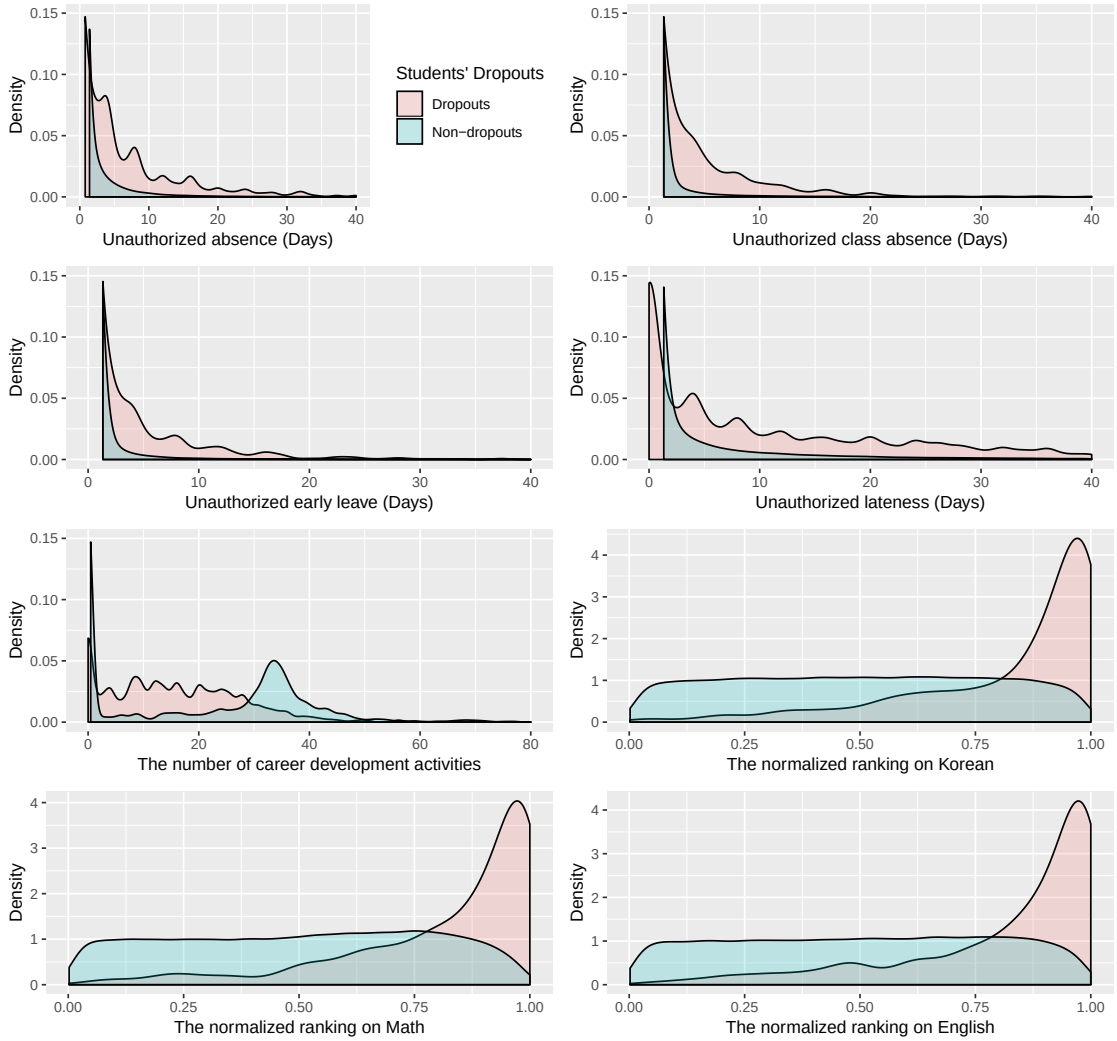

Figure 1. The density plots for the selected eight features.

Figure 2 presents the ROC curves for the four binary classifiers used in this study. The AUC of the random forest (RF), boosted decision tree (BDT), random forest with SMOTE (SMOTE + RF), and boosted decision tree with SMOTE (SMOTE + BDT) were 0.986, 0.988, 0.986, and 0.991, respectively.

Figure 3 presents the PR curves for the four binary classifiers used in this study. The AUC of the random forest (RF), boosted decision tree (BDT), random forest with SMOTE (SMOTE + RF), and boosted decision tree with SMOTE (SMOTE + BDT) were 0.634, 0.898, 0.643, and 0.724, respectively.

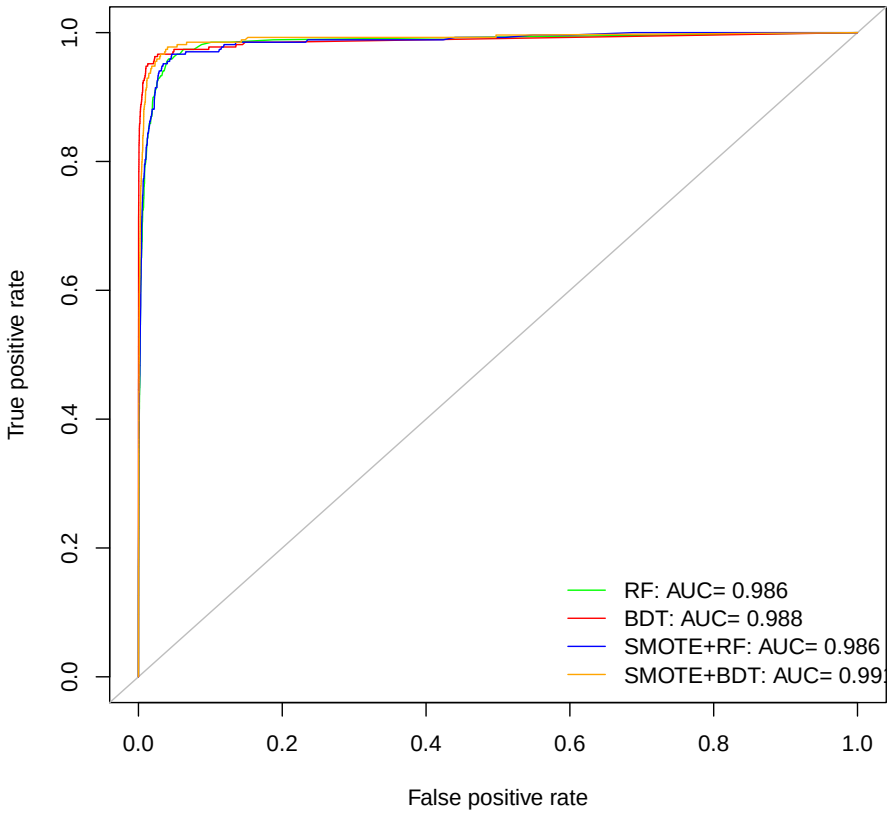

Figure 2. The ROC curves.

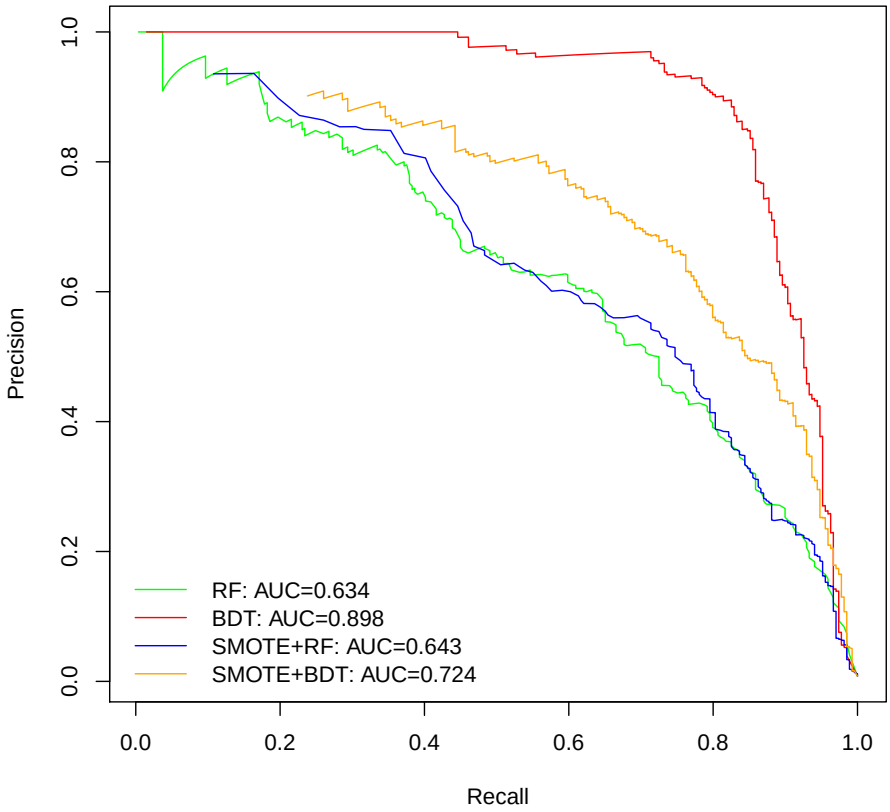

Figure 3. The PR curves.

6. Discussion

Classifiers trained on class-imbalanced datasets tend to show a poor sensitivity of predicting minor classes because classifiers tend to ignore the misclassification of minor classes. Given our specific goal of maximizing our chance of supporting the successful learning of all students, and minimizing the cost for intervention, the present study aimed to improve the performance of a dropout early warning system: (a) by addressing the class imbalance issue using the SMOTE and ensemble methods in machine learning; and (b) by evaluating the trained classifiers with both receiver operating characteristic (ROC) curves and precision–recall (PR) curves. Using the reliable features from the big data samples of the 165,715 high school students provided by the NEIS database in Korea, we trained four classifiers: random forest (RF), boosted decision tree (BDT), random forest with SMOTE (SMOTE + RF), and boosted decision tree with SMOTE (SMOTE + BDT). Based on our model-wide evaluation based on ROC and PR curves, boosted decision tree showed the best performance.

In Figure 2, the four ROC curves indicate that all four models were excellent in terms of AUCs: the AUC values of RF, BDT, SMOTE + RF, and SMOTE + BDT were 0.986, 0.988, 0.986, and 0.990, respectively. However, this result needs to be interpreted with caution. As previously discussed, the ROC curve may not be a good evaluation metric for imbalanced datasets because the false positive rate (or $1 - \text{specificity}$ or $1 - \text{TN}/(\text{FP} + \text{TN})$) does not change much when the true negative (TN) is huge, which is common for imbalanced datasets. In our analysis, the ROC curves were not informative when comparing the performance of our four classifiers. On the contrary, the PR curves in Figure 3 were more informative in that the PR curves and their corresponding AUC values were more distinctive: the AUC values of RF, BDT, SMOTE + RF, and SMOTE + BDT were 0.634, 0.898, 0.643, and 0.724, respectively. According to the AUC values of the PR curves, the BDT showed the best performance (i.e., AUC = 0.898), indicating that, among the four tested classifiers, the dropout early warning system based on BDT was optimal in maximizing our chance of supporting the successful learning of all students, and minimizing the cost for intervention. Our result is consistent with the recent study on the impact of class rebalancing techniques on the performance of prediction models. Tantithamthavorn et al. [34] recently found that class rebalancing techniques, such as SMOTE, impact recall the most positively and impact precision the most negatively.

The PR curve essentially illustrates the trade-off between recall (or sensitivity) and precision. This trade-off between recall and precision raises an important issue regarding decision making based on machine learning. The predictions made by machine learning are often used to make a policy or plan, which always has a budget constraint. Therefore, the optimal performance of a classifier is only meaningful in relation to a specific task with the consideration of both the benefit from the improved performance and the cost for the improvement. In this study, our task was to build the dropout early warning system that maximizes the sensitivity of predicting potential dropout students to support successful learning of all students, and minimizes the cost for intervention. Even with BDT, the false alarms still exist and will require additional costs in time and money for the interventions. However, we believe that the benefit of preventing dropout students exceeds the cost for interventions. For example, in Ohio in the United States, the median earnings of a high school dropout are $17,748, whereas the ones of high school graduates are $26,207 [35]. The additional earning of $8,459 a year would be accumulated over a lifetime of a single individual.

The results of this study should be interpreted with caution because of the different nature of performance metrics. Sensitivity and specificity are independent of the prevalence of positives in the population, whereas positive and negative predictive values are influenced by the prevalence of positives in the population. Therefore, if our predictive modeling were transported to a population with higher frequencies of dropouts, the sensitivity would remain the same because the sensitivity is the characteristic of a test, but the positive predictive value would increase because the positive predictive value reflects the population.

In machine learning, the quality of training data is a critical factor that determines the performance of predictive models. The NEIS is an ideal database for developing the dropout early warning

system in South Korea for several reasons. First, the NEIS database contains information about more than six million students' basic information, academic achievement, absence, health and so on. Therefore, the NEIS can provide various students' features to build an effective dropout early warning system. Second, the NEIS has been used during the past 20 years in South Korea. Therefore, teachers are well trained in using the NEIS system, which also improves the quality of data.

There are some limitations to the present study. First, our access to the NEIS database was limited in this study. Although we included the key risk indicators for dropout prediction in our analysis, we were not able to access many other features in the NEIS database, such as teachers' evaluation of students, at the time of our analysis. We expect that the performance of our prediction model could be improved by adding those additional features in the future. Second, our predictive model only predicts students at risk of dropping out of school. Recent advances in predictive modeling enable us to estimate heterogeneous treatment effects to understand those who can effectively be intervened [36]. Such information would be very useful in designing and implementing prevention programs.

In sum, we aimed to build a dropout early warning system that maximizes our chance of supporting the successful learning of all students, and minimizes the cost for intervention by addressing the class imbalance issue using the SMOTE and ensemble methods in machine learning, and also by evaluating the trained classifiers with both ROC and PR curves. ROC curves were not very informative, whereas PR curves were informative. According to our PR curves, BDT showed the best performance. Considering the prevalence of the class-imbalance issue in other educational outcomes, this study has implications for other educational studies using predictive modeling as well.

Author Contributions: Conceptualization, S.L. and J.Y.C.; methodology, S.L.; software, S.L.; validation, S.L. and J.Y.C.; formal analysis, S.L.; investigation, S.L. and J.Y.C.; resources, S.L. and J.Y.C.; data curation, J.Y.C.; writing—original draft preparation, S.L.; writing—review and editing, S.L. and J.Y.C.; visualization, S.L.; supervision, J.Y.C.; and project administration, J.Y.C.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rumberger, R.W. High School Dropouts: A Review of Issues and Evidence. *Rev. Educ. Res.* **1987**, *57*, 101. [CrossRef]
2. Catterall, J.S. On the social costs of dropping out of school. *High School J.* **1987**, *71*, 19–30.
3. Balfanz, R.; Herzog, L.; Maciver, D.J. Preventing Student Disengagement and Keeping Students on the Graduation Path in Urban Middle-Grades Schools: Early Identification and Effective Interventions. *Educ. Psychol.* **2007**, *42*, 223–235. [CrossRef]
4. Dynarski, M.; Gleason, P. How Can We Help? What We Have Learned From Recent Federal Dropout Prevention Evaluations. *J. Educ. Stud. Placed Risk (JESPAR)* **2002**, *7*, 43–69. [CrossRef]
5. Lamb, S.; Rice, S. *Effective Strategies to Increase School Completion Report: Report to the Victorian Department of Education and Early Childhood Development*; Communications Division, Department of Education and Early Childhood Development: Melbourne, Australia, 2008.
6. Knowles, J.E. Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *JEDM J. Educ. Data Min.* **2015**, *7*, 18–67.
7. Sullivan, R. *Early Warning Signs. A Solution-Finding Report*; Center on Innovations in Learning, Temple University: Philadelphia, PA, USA, 2017.
8. Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **2017**, *73*, 220–239. [CrossRef]
9. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Trans. Syst. Man Cybern. Part C* **2012**, *42*, 463–484. [CrossRef]
10. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]

11. Márquez-Vera, C.; Cano, A.; Romero, C.; Ventura, S. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Appl. Intell.* **2013**, *38*, 315–330. [CrossRef]
12. Márquez-Vera, C.; Cano, A.; Romero, C.; Noaman, A.Y.M.; Mousa Fardoun, H.; Ventura, S. Early dropout prediction using data mining: A case study with high school students. *Expert Syst.* **2016**, *33*, 107–124. [CrossRef]
13. Korean Educational Development Institute. Available online: http://cesi.kedi.re.kr/post/6662567?itemCode=03&menuId=m_02_03_03 (accessed on 1 August 2018).
14. Chung, J.; Kang, T.; Kim, S.K.; Ryoo, J.S.; Lee, D.; Lee, J.; Hwang, J. *Policy Study on the Supporting System for Out-of-School Youth*; Jeollanamdo Office of Education: Jeollanamdo, Korea, 2013.
15. Peters, S.J. "Education for All?". *J. Disabil. Policy Stud.* **2007**, *18*, 98–108. [CrossRef]
16. Mathis, W.J.; Trujillo, T.M. *Lessons from NCLB for the Every Student Succeeds Act*; National Education Policy Center: Boulder, CO, USA, 2016.
17. Chung, J.Y.; Lee, S. Dropout early warning systems for high school students using machine learning. *Child. Youth Serv. Rev.* **2019**, *96*, 346–353. [CrossRef]
18. Chawla, N.V.; Lazarevic, A.; Hall, L.O.; Bowyer, K.W. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In *Knowledge Discovery in Databases: PKDD 2003*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 107–119. [CrossRef]
19. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
20. Hand, D.J.; Whitrow, C.; Adams, N.M.; Juszczak, P.; Weston, D. Performance criteria for plastic card fraud detection tools. *J. Oper. Res. Soc.* **2008**, *59*, 956–962. [CrossRef]
21. Chawla, N.V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 853–867. [CrossRef]
22. Longadge, R.; Dongre, S. Class imbalance problem in data mining review. *arXiv* **2013**, arXiv:1305.1707.
23. Lin, Y.; Lee, Y.; Wahba, G. Support vector machines for classification in nonstandard situations. *Mach. Learn.* **2002**, *46*, 191–202. [CrossRef]
24. Napierała, K.; Stefanowski, J.; Wilk, S. Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In *Rough Sets and Current Trends in Computing*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 158–167. [CrossRef]
25. Ling, C.; Sheng, V.; Yang, Q. Test strategies for cost-sensitive decision trees. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1055–1067. [CrossRef]
26. Zhou, Z.H. *Ensemble Methods*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2012. [CrossRef]
27. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001. [CrossRef]
28. Zhou, Z.H.; Jiang, Y.; Yang, Y.B.; Chen, S.F. Lung cancer cell identification based on artificial neural network ensembles. *Artif. Intell. Med.* **2002**, *24*, 25–36. [CrossRef]
29. Panigrahi, S.; Kundu, A.; Sural, S.; Majumdar, A. Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Inf. Fusion* **2009**, *10*, 354–363. [CrossRef]
30. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
31. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
32. Therriault, S.B.; Heppen, J.; O'Cummings, M.; Fryer, L.; Johnson, A. Early Warning System Implementation Guide. Retrieved from the National High School Center Website. 2010. Available online: http://www.betterhighschools.org/documents/NHSCEWSImplementationGuide.pdf (accessed on 1 August 2018).
33. Allensworth, E.M.; Easton, J.Q. *What Matters for Staying On-Track and Graduating in Chicago Public High Schools: A Close Look at Course Grades, Failures, and Attendance in the Freshman Year. Research Report*; Consortium on Chicago School Research: Chicago, IL, USA, 2007.
34. Tantithamthavorn, C.; Hassan, A.E.; Matsumoto, K. The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models. *IEEE Trans. Softw. Eng.* **2018**, 1. [CrossRef]

35. Cellini, S.R.; Kee, J.E. Cost-Effectiveness and Cost-Benefit Analysis. In *Handbook of Practical Program Evaluation*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2015; pp. 636–672. [CrossRef]
36. Wager, S.; Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J. Am. Stat. Assoc.* **2018**, *113*, 1228–1242. [CrossRef]