Applying a Hybrid Sampling and Boosting Approach to Predict Student Retention

Eric P. Jiang

***Abstract*—Over the past decades, machine learning has been successfully applied to every sector of business to help learn customer needs, make intelligent decisions, and to better serve customers. This includes institutions of higher educations. Specifically, machine learning models can be built on the data from educational settings and used to improve student learning experiences as well as institutional effectiveness. In this paper we propose an innovative approach for solving class imbalanced problems and apply it in student retention prediction. The approach employs a newly developed hybrid data sampling procedure and boosting algorithm to enhance classification performance on data with imbalanced class distribution. Experiments with a collected student data set indicate that the proposed approach is capable of classifying data with limited info and skewed class distribution effectively and furthermore, in comparison with several popular learning algorithms that include decision trees, naïve Bayes and support vector machines, their cost-sensitive counterparts, as well as RUSBoost and SMOTEBoost, it delivers a superior classification performance.**

***Index Terms*—Imbalanced classification, data sampling, ensemble algorithms, student retention.**

I. Introduction

Over the past decades, machine learning has been successfully used in every sector of business to learn consumer needs, make intelligent business decisions and to better serve customers. Institutions of higher education have also started on a similar path to use their knowledge about students to enhance student services and institutional effectiveness [1]. With the advancement of machine learning technology, universities can do a much better job in providing supportive student learning environment and improving student learning experiences, for instance, by providing effective support and intervention services, throughout a semester, to the students who are needed most to succeed than just waiting to the end of the semester to tell those students that they are not doing well in school [2].

Among several strategic objectives that most institutions of higher education attempt to achieve, improving and maintaining a high student retention rate is a very important one. The success of an institution is measured by the success of its students. To a large extent, this is reflected by student retention and graduation in a timely manner as well as student preparation for the workforce and citizenship. As a transition from high school to college, it can be quite easy for students, especially in their early semesters in higher education, to get lost in a new learning and living environment. For some students, this can be due to their lack of self-discipline and/or deficit in preparation for higher education. For instance, the data from the National Center for Public Policy and Higher Education reveal that only 73.6% of full time freshmen enrolled in 2002 returned for their second semester. In addition, looking at college completion data between 2005 and 2010, only 39.5% of undergraduate students enrolled in U.S. public institutions completed their degrees within five years [3].

There are a number of negative consequences associated with a low student retention rate. First, institutions would face a loss in revenue for each of dropouts. More dropouts mean fewer graduates, which can further lead to fewer alumni and potentially fewer gifts in the future. Second, student retention also has an impact in racial and ethnic diversity among college students as this has been a more noticeable issue among students of color. Third, both the retention and graduation rates make up a considerable portion of the popular US News university ranking score (with 20% and 5%, respectively) [1]. In order to provide a supportive learning environment that fosters student success and retains students, institutions need to identify and understand important factors that may impact student retention. Furthermore, these factors can be utilized to build machine learning models to predict which students are potentially at risk of dropout. As aforementioned, being able to identify these factors and individuals with respect to retention will certainly help institutions offer effective and targeted support and intervention services to those who need most to succeed.

There are a number of negative consequences associated with a low student retention rate. First, institutions would face a loss in revenue for each of dropouts. More dropouts mean fewer graduates, which can further lead to fewer alumni and potentially fewer gifts in the future. Second, student retention also has an impact in racial and ethnic diversity among college students as this has been a more noticeable issue among students of color. Third, both the retention and graduation rates make up a considerable portion of the popular US News university ranking score (with 20% and 5%, respectively) [1]. In order to provide a supportive learning environment that fosters student success and retains students, institutions need to identify and understand important factors that may impact student retention. Furthermore, these factors can be utilized to build machine learning models to predict which students are potentially at risk of dropout. As aforementioned, being able to identify these factors and individuals with respect to retention will certainly help institutions offer effective and targeted support and intervention services to those who need most to succeed.

Manuscript received July 4, 2021; revised March 11, 2022. .

E. P. Jiang is with the University of San Diego, San Diego, CA 92110 USA (e-mail: jiang@sandiego.edu).

doi: 10.18178/ijmlc.2022.12.5.1099

II. Background

Over the years there has been a considerable amount of research efforts that address the student retention issue. The early theoretical models of student retention can be dated back several decades. The popular Tinto's model [4] investigates factors associated with student decisions to continue their education with the institution. The model primarily focuses on a student's academic and social integration into the institution and it implies that students tend to remain with their programs when they perceive their institution is a suitable place to achieve their academic goals [2]. There are several alternative retention models that, for instance, focuses on other contributing factors such as in psychological and behavioral aspects [5].

Student retention rates are generally calculated based on data from first-time, full-time freshmen who graduate within six years of their initial enrollment date [6]. Among all academic years, freshman year represents a stressful transition for university students. Despite a multitude of social, academic and emotional issues, most students successfully cope with a new and complex college life and achieve their academic success. But some other students are less able to manage this transition adequately and decide to leave their school during or at the end of their freshman year [7]. As freshman class attrition rates are generally higher than other classes and the intervention program can have more significant effects on retention for the first year [8][9], there is a good portion of retentions studies that devote to issues and possible ways to improve freshman retention (e.g., [10][11]) while several others focus on student persistence beyond the freshman year (e.g., from sophomore to junior) [3] [12]. This paper, however, addresses a quite unique issue, namely student retention after the first semester, and we are unaware of any similar research work in literature.

In terms of popular modeling techniques used in this application domain, particularly in the area of retention and graduation of university students, they have transformed over the years. Specifically, they have moved from traditional parametric statistical approaches such as regression analytics and logistic modelling to more powerful machine learning methodologies. Indeed, machine learning models generally work notably better with a large number of predictors and can more effectively capture nonlinear relationships and complex interactions among predictors as well as between predictors and the target variable. In particular, decision trees and artificial neural networks have been the two popular choices in this research area.

Student retention prediction based on modeling is a very challenging classification problem. Like many other applications such as fraud detection and medical diagnosis, we have to deal with the class imbalance issue in model building. More specifically, in a typical institution setting, the number of retained students (as the majority class) significantly outnumbers the number of dropouts (as the minority class). Traditional classification algorithms generally fail to work adequately with skewed class distribution problems as they are designed to generalize from sample data and produce the simplest hypothesis that best fits the data. Clearly, such a hypothesis can simply be worthless in practice.

A number of approaches have been proposed to address the challenges of imbalanced classification [13][14]. Some of the approaches are at algorithm level, by creating new algorithms or adapting the existing ones to shift an inductive bias towards the minority class. A popular adapting methodology to mitigate class imbalance takes the misclassification costs into account to build cost-sensitive classifiers.

In traditional learning, we treat all misclassifications equally. But this can cause issues with imbalanced datasets as it tends to create learning models that are biased towards classifying the majority class over the minority class. In many real-world applications such as student retention, we are more interested in identifying instances of the minority or dropout class. We assume the minority class as the positive and the majority as the negative and consider a function C(i, j) that specifies the cost of misclassifying an instance of class i as class j. As the recognition of positive instances is generally more important than that of negative instances, C(+, -) > C(-, +). With the cost function, cost-sensitive learning algorithms seek to minimize the number of expensive errors and the total misclassification cost (rather than maximizing the percentage of accuracy). As for the misclassification cost of a class, we may or may not know the value in practice. In this case, a popular scheme is to set the cost equal to the inverse of the proportion of the dataset that the class makes up. The scheme is used in our experiment.

Another group of the solutions to strengthen the learning with respect to the minority class are at data level and more specifically, they resample the data space to reduce the potentially negative effect of class imbalance on model building. Under-sampling the majority class or over-sampling the minority class are the popular choices in this group. However, in practice, under-sampling may cause information loss for the majority class while over-sampling may lead to model overfitting as it duplicates instances from the minority class that is already small. There also exist more complex sampling methods beyond simple over-sampling and under-sampling. One of the best-known algorithms in this category is SMOTE [15] in which the minority class is over-sampled by creating new synthetic samples along the line segments joining neighboring minority instances. More recently, several different approaches that combine sampling techniques and ensample algorithms such as boosting and bagging have been proposed as solutions to class imbalance problems [13].

III. The Student Retention Problem and the Proposed Approach

A. The Freshman Spring Retention Problem

As it is aforementioned, in comparison to other student classes, freshmen are generally more at risk of drop-out. Furthermore, among those freshman drop-outs, a good portion of them leave their schools just after the first fall semester (do not return and enroll in the following spring semester). We call this the freshman spring retention problem.

For many freshmen, their very first semester in college can be the most stressful one as they jump into a new living and learning environment. As a result of this, the attrition rates for freshmen after the first semester can be quite noticeable or even significant, depending upon school types. For some

private colleges with a small campus and class size, the rates might be around 5% while for some public universities with a large campus the rates could be even beyond 25%. Being able to identify who are at risk of dropping out during an early stage could help universities allocate intervention resources such as advising and mentoring programs to the right students at the right time and consequently help retain some of the at-risk students. In addition, it can be argued that reducing the initial attrition rate during the freshman year has another advantage and it can help universities learn more about their students and develop more effective and practical strategies and processes to further improve student retention in subsequent semesters or years.

The freshman spring retention is a very challenging problem to deal with. Frist, we generally don't have adequate and sufficient data or information about students to develop powerful machine learning solutions. Student retention indeed is a very complex issue itself and there can be so many possible contributing factors related to retention and they can even be quite different from one student to another. Some contributing factors may include the lack of academic preparation and low confidence level in students, the lack of an appropriate peer community and an alienation from the environment and possibly the lack of adequate financial aid. The lack of relevant student info is particularly severe for predicting freshman spring retention. As a result of this, we can only use the standard pre-college student data typically collected and used in admission processes. On the other hand, this limited data source used for profiling at-risk students could mean at the same time that any reasonably working solutions to this special retention problem would be applicable to almost all institutions as they do not require additional and specially gathered student data that can likely be unique to a particular institution.

The other major difficulty to deal with this problem lies in skewed class distributions in student data. Generally, as an imbalanced classification problem, the number of retained freshmen in the second semester outnumbers the number of dropouts and the ratio between the retained and dropout classes can even be quite drastically for many institutions. As discussed in Section II, there are a number of approaches have been proposed to address imbalanced classification but it seems that they do not work well in this special data mining application. We will discuss this in further detail in the following subsection and Section IV. Furthermore, the limited data used for freshman spring retention are also generally very noisy in the sense that there is no clear boundary between the two classes. In other words, a good portion of freshmen, although belong to different classes, may share some common characteristics. Several factors, reckoned as relevant to retention by previous publications in literature, such as changes of family finance and student academic adjustment to college are usually unknown or unavailable by the time for student spring retention decisions.

B. The Proposed Approach

We propose a new approach that integrates a two-stage data under-sampling strategy and the standard boosting algorithm and it aims to effectively deal with data that are both imbalanced and noisy. In this paper we use the approach for a binary classification problem (student retention prediction).

Since freshman retention identification is a highly imbalanced classification problem, the very skewed class distributions of the retained and the dropouts can make a typical classification algorithm to heavily rely on the training sample of the retained freshman (majority) class. Consequently, it can produce inaccurate predictions of the dropout (minority) class, which leads an extremely low false positive rate for the class. One of the popular approaches to work with such skewed data distributions is to create a more balanced training data set between the classes by either randomly under-sampling the majority class or alternatively over-sampling the minority class through duplicates. However, these simple data sampling strategies would only be helpful for handing moderate class imbalances. In more extreme imbalance problems such as freshman spring retention, one possible solution is to consider the problems in the context of anomaly detention. Specifically, we assume there is a normal distribution among the data from the majority class and by applying anomaly detection algorithms, we can search for anomalies that sufficiently deviate from the normal distribution and subsequently classify them to the minority class. Of course, the assumption of a normal distribution for the majority class may not always be applicable to real-world data in practice, which is particularly the case when the data are noisy. Thus, a simple execution of this approach can still lead to significant misclassifications.

In the rest of the subsection, we describe a two-stage approach that integrates a unique data sampling procedure with the standard boosting algorithm and it aims to effectively address the issue of imbalanced data distribution and reinforce the learning performance. The approach has been used to predict the freshman spring dropouts with relatively high true positives.

In the first stage of the approach, we form a balanced training data set through an under-sampling procedure. This is done by applying an anomaly detection algorithm on all given majority (or retained student) data samples to separate them whose behaviors might be common among the majority population from others whose behaviors might be somewhat less typical from the same population. The objective of this stage is to create a balanced and also well-representative training dataset for the majority class, in particular in the presence of data noises and class overlapping regions. A similar strategy has recently been applied to identify thieves in public transit systems [16].

There are a number of unsupervised anomaly detection algorithms that can be used to identify outliers. For this approach, we use the one class support vector machines (SVM) algorithm [17] because of its solid theoretical foundation, efficient computations and superior classification performance. Like the standard SVM, one class SVM computes non-linear decision boundaries by using appropriate kernel functions and soft margins but it constructs the decision boundaries that separate the majority of the data from the origin. Only a portion of the data points are allowed to be on the other side of the boundaries and these pointes are regarded as outliers.

More specifically, assuming we have a set of training data with *n* samples

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \quad (1)$$

where x_i are the vectors in some space X representing samples' attribute values and $y_i \in [+1, -1]$ are the samples' class labels. Using a kernel function defined by $\kappa(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, where the function $\phi(\cdot)$ transforms the data points x_i from X to a high dimensional feature space F, the optimal decision boundary in the transformed space F can be expressed as

$$w^T \phi(x_i) - \rho = 0 \quad (2)$$

where w is the vector in F perpendicular to the decision boundary and ρ is the bias term. The optimal decision boundary for one class SVM can then be obtained by solving the following constrained optimization problem

$$min_{w,\rho,\xi_i} \frac{\| w \|^2}{2} + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i - \rho$$

$$subj\ to\ w^T \phi(x_i) \geq \rho - \xi_i,\ \xi_i \geq 0,\ i = 1, 2, \cdots, n \quad (3)$$

where ξ_i is the slack variable to allow x_i to be lie on the other side of the decision boundary and ν is the regularization parameter.

By using a kernel function, we can find the desired decision boundary without dealing with the exact form of the transformation function $\phi(\cdot)$. We use the well-known Gaussian kernel for our one class SVM implementation as it guarantees the existence of such an optimal decision boundary separating data from the origin [17].

Now, for a given set of training data, we apply the one class SVM algorithm to all samples of the majority or retained class to partition them into the regular portion that has the samples sharing some mainstream behaviors and the anomalous portion that has the samples having relatively unique behaviors. We can manipulate the regularization parameter to have a desired data split between the two portions. For instance, we can choose a portion of data like 20% or 30% as anomalous and the rest of them as regular for the partition. Then, according to our modelling needs, we randomly select a certain percentage of the majority or retained from the regular portion and the remaining from the anomalous portion to generate a final set of training data of the majority samples or retained freshmen that are comparable in size with that of dropouts. In the experiments discussed in Section 5, we formed the retained freshman set by selecting 40% and 60% of them from the regular portion and the anomalous portion, respectively.

Once a balanced training dataset is formed, we move to the second stage of the approach. In this stage, we use the standard boosting algorithm, AdaBoost, to build a classification model. AdaBoost, or Adaptive Boosting, is a well-known ensemble machine learning algorithm [18] and it can be used in conjunction with other base learning algorithms to improve classification performance. The method involves an iterative process that produces a sequence of classifiers or hypotheses and, in each step of the process, it builds a classifier that focuses more on the training samples that are misclassified by the previous one. This is accomplished by using an adaptive weighting scheme on the training data. Specifically, the boosting procedure takes as input a training dataset, say S (1), and applies a learning algorithm repeatedly in multiple rounds to builds an ensemble of classifiers. It begins with an initial distribution D_1 of S by assigning an equal weight w_i to all training samples,

$$D_1(i) = w_i = 1/n, \quad i = 1, 2, \ldots, n$$

and on round t, it applies the algorithm to build a classifier or hypothesis, as

$$h_t(x): X \rightarrow \{-1, +1\}. \quad (4)$$

Then, the procedure computes the classification error e_t of $h_t(x)$ with respect to the distribution D_t

$$e_t = \sum_{i:\, h_t(x_i) \neq y_i} D_t(i)$$

and the sample weight update term

$$\beta_t = e_t / (1 - e_t) \quad (5)$$

which is used to update D_t and to produce the final hypothesis. The distribution of S for the next round, D_{t+1}, is updated so that the weights are multiplied by β_t (5) for all correctly classified training samples by $h_t(x)$(4),

$$D_{t+1}(i) = D_t(i) \times \beta_t, \quad i = 1, 2, \ldots, n$$

and there are no weight changes for the misclassified samples. Once all sample weights are updated, they are normalized and this normalization step effectively increases the weight for misclassified samples and decreases the weight for correctly classified ones. The process is repeated for a number of times to generate a sequence of classifiers. The final classification output from the procedure is formed by a weighted majority vote of the models

$$h_{final}(x) = arg\ max_y \sum_t h_t(x) \times log \frac{1}{\beta_t} \quad (6)$$

where the weight for the individual model is given by log $(1/\beta_t)$. This weight scheme gives a high weight to a model that performs well on the training samples and gives a low weight to a model that performs poorly.

It is noted that the adaptively adjusted weighting structure of AdaBoost can actually help alleviate the imbalance data distribution problem we have with freshman retention, which is an important reason for us to choose it as our second stage learning algorithm. As described above, if the minority or dropout class cannot be learned very successfully at early iterations, the system automatically adds more weight to the corresponding training samples and tries to learn them correctly at subsequent iterations. Therefore, the proposed approach with AdaBoost has the potential to help deliver good classification results for the freshman retention prediction problem.

IV. Several Popular Classification Approaches for Comparison

In this section, we describe several classification algorithms we have selected to compare with the proposed two-stage approach. The algorithms can be grouped into 3 different sets. The first set of the algorithms are traditional classifiers and particularly well-known in the educational data mining community. It includes decision trees (DT) [19], Naïve Bayes (NB) [20], logistic regression (LR) [21] and Support Vector Machines (SVM) [22], [23]. Although these classifiers are widely used in a variety of classification applications, they are not designed for handling problems

with imbalanced class distributions directly. By selecting this set of algorithms for the freshman retention problem and evaluating the obtained results, it will help reveal how challenging this problem is as well as how important to develop more adequate learning solutions to it.

The second set of the algorithms we selected for comparison are the cost-sensitive versions of the first set. As described in Section II, they represent one of the primary solutions to class imbalance problems. There are several ways to implement cost-sensitive learning. One approach is to build a classifier without using the costs and only apply them at prediction time to adjust the classification threshold. A better alternative is to assign the training samples of different classes with different weights, which are in proportion to their costs, and then use the weighted samples to build a classifier [24]. With this arrangement, the classifier can place more emphasis on samples with higher weights during the training process and for class imbalance problems, this could lead to improved identification of the minority class. We used the latter approach for the algorithms in this set.

The last set of the classifiers included in our experiment is to use some resampling strategies. They represent another primary consideration for developing effective solutions to mitigate class imbalance. As we have discussed in Section 2, balancing an imbalanced dataset can be accomplished by either over-sampling instances of the minority class or under-sampling the majority class. Sampling can be done randomly or by some more sophisticated schemes such as SMOTE. Extensive experiments reported in [13] have shown, among different solutions to imbalance classification, the approaches that combines random under-sampling (RUS) and SMOTE with boosting are very promising. In this paper, we include two of the best-known combination algorithms, RUSBoost [25] and SMOTEBoost [26], into our third set of algorithms to be compared with.

V. Experiments

A. Data and Data Preprocessing

The data used in this project was acquired from a liberal arts university located on the west coast of the United States of America. The original raw data, directly pulled out from the school's applicant database from 2002 to 2009, contains over 50 attributes that include student id, age, gender, high school GPA, SAT scores, financial aid (scholarships), and most of them related to student demographics and academic preparation. The data set contains 8,959 freshman records for those enrolled in the university between 2002 and 2009 and among all enrolled, 424 freshmen left the school after their first spring semester and the rest 8,535 freshmen continued their study at the school after that semester. As the dropout class takes a very small percentage (4.73%) of the population, the dataset is highly imbalanced in class distribution.

Like many other real-world data, this freshman collection contains numerous missing entries. For instance, there is an interesting attribute called high school percentile in the original set that can bring some value for model learning. But unfortunately, the attribute contains too many missing entries to make it useful and we had to remove it from the set. On the given data, we also performed a number of comprehensive data preprocessing tasks that include missing value substitution, comparable value conversion and outlier detection and handling. In addition, we took several random samples from the dataset to learn their attribute distributions and, combining with previous student retention research work in the literature (e.g., [27]), we selected a total of 15 attributes including the target (Retained), as shown in Table I.

TABLE I. The List of Selected Attributes

SAT/ACT scores	High school GPA	Catholic high school
Gender	Domestic	State resident
Distance to campus	Financial aid $	State grant recipient
Honors program	Live on campus	Merit scholarship recipient
African American	Catholic	Retained

B. Experiment Setup

As we described in the previous section, the proposed two-stage approach integrates a new under-sampling procedure and the standard boosting algorithm (AdaBoost) to produce a competitive solution to class imbalance problems. The under-sampling procedure in the first stage applies one class SVM to partition the data of the majority class into two portions: one with the data points that share some common characteristics and the other that behave relatively in their individual ways. In this study, once the portions were produced, we formed a training set of the majority or retained class by randomly selecting an equal share from both portions and the resulting dataset has a comparable size to the original minority or dropout class. As another under-sampling scheme, RUSBoost formed the training dataset in a similar manner but it simply selects the samples randomly. For SMOTEBoost, in order to fully utilize original training samples, we first created a new set of synthetic minority instances and, together with the original minority instances, it has a total that is half of the given majority class. Then we randomly selected the same amount of majority class and combined it with the selected minority set to form the final training dataset. In the second stage of the proposed approach, we apply AdaBoost on the balanced training samples, in hopes that it further reduces the bias of the final classifications along with the variance to improve classification performance.

There are several metrics that can be used for assessing classification performance and guiding model learning. The accuracy is the most popular metric for general classification. However, it is not adequate for evaluating solutions to class imbalance problems as the minority class has little impact on the metric compared to the majority class. In other words, in these cases accuracy reveals more about distribution of classes than it does about the actual performance of models. Freshman spring retention is a class imbalanced problem and typically there are only a very small percentage of the students who leave the school after the first semester.

Recall or sensitivity is a widely used metric for the class imbalance problems where successful prediction of the minority class (e.g., dropout) is considered more significant and useful than prediction of the opposite majority class (e.g., retained). In our case with freshman spring retention, sensitivity is the percentage of freshman dropouts are correctly predicted by the classifier and therefore, it assesses the classifier's effectiveness on the dropout class. There is a corresponding metric, called specificity, to measure the

accuracy of the retained class. It is the percentage of retained freshmen correctly predicted by the classifier.

Clearly, there is a trade-off between sensitivity and specificity values but for student retention, we are usually more interested in detecting the dropout class and concerned more so with sensitivity than specificity, as it likely would be costlier to miss a dropout than to falsely flag a retained student. In other words, it should be relatively doable to verify that a student is actually staying with the school in the latter case, but it would be much harder to identify dropouts that were never labeled as such.

The geometric mean (G-mean) of sensitivity and specificity has also been a popular unified metric for imbalance class applications. It is defined as

$$G - mean = \sqrt{sensitivity \times specificity}$$

and can be used to measure the balance of classification performances on both classes and in particular to help determine if a classifier overfits the majority class and underfits the minority class. A low G-mean value would typically indicate a poor performance in the classification of the minority class even if the majority class is mostly correctly classified.

Another popular and useful tool for evaluating and comparing performance between different classifiers for class imbalance is the receiver operating characteristics (ROC) curve. For a given classifier, the ROC curve plots the true positive rate of the classifiers on the vertical axis against its false positive rate on the horizontal axis. To compare different models with ROC, however, it is hard to declare a winner unless one curve dominates the other(s) over the entire space. As an alternative, the area under a ROC curve (AUC) provides a singer performance measure for evaluating and determining which model is better on average. We used sensitivity, AUC and G-mean in our classifier evaluation and comparison.

All experiments in this study were conducted through the 5-fold cross validation process. The results reported in the following subsection are the averaged classification performance metrics.

C. Experiment Results

Fig. 1, Fig. 2 and Fig. 3 shows the classification results in recall or sensitivity, AUC and G-mean, respectively, from the proposed two-stage approach and also from several other popularly used classification algorithms when applied to detect potential freshman spring dropouts. The figures can also be used as a direct comparison between these difference approaches.

We would like to offer a few remarks for the results shown in the figures. First, it is somewhat expected that the traditional classification algorithms (the first four classifiers in the figures) that are designed to maximize accuracy fail to work for this freshman spring retention problem and this is reflected clearly by their undesirable recall or sensitivity values. In particular, both naïve Bayes and SVM are not be able to identify any dropouts. This indicates the limitation of this type of algorithms when applied to class imbalance data sets.

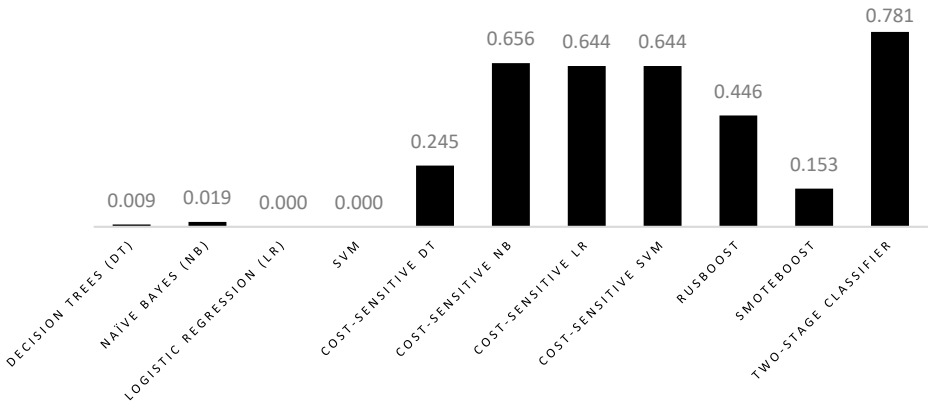

Fig. 1. The recall or sensitivity comparison between the proposed classifier and other approaches.

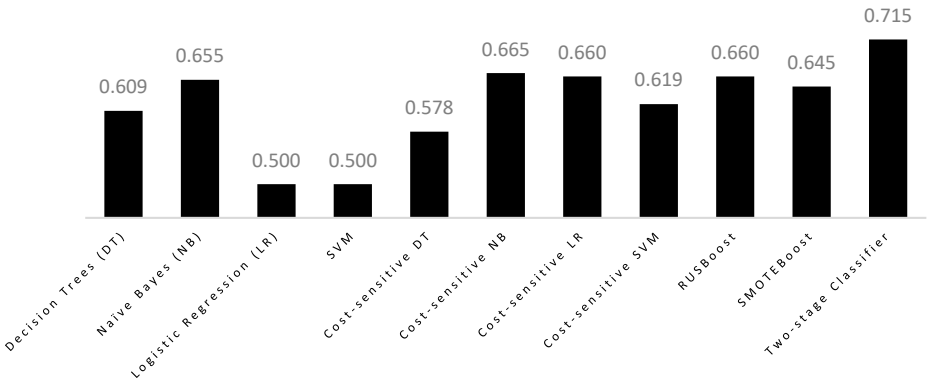

Fig. 2. The AUC comparison between the proposed classifier and other approaches.

Second, when these algorithms are modified to take misclassification costs of different classes into consideration in model learning, they deliver significantly better classification performance, especially for naïve Bayes,

logistic regression and SVM. These are shown by the middle four algorithms in the figures. Among them, Cost-sensitive DT is underperformed by other three cost-sensitive counterparts (namely, cost-sensitive NB, cost-sensitive LR and cost-sensitive SVM) that behave very comparably each other in all considered metrics.

Third, comparing with most of cost-sensitive approaches considered in the experiment, the well known RUSBoost and SMOTEBoost actually deliver an inferior recall or sensitivity value for the minority class. This is particularly the case for SMOTEBoost. As an over-sampling approach, SMOTEBoost builds the learning model based on both original majority class instances and synthetically expanded minority class instances and it seems this helps shift its modeled predictions toward the majority class, which leads to a low recall for the minority class.

Finally, comparing all other algorithms included in this experiment, the proposed two-stage approach delivers an overall superior and more balanced performance for freshman retention prediction. This approach under-samples the majority class based on its partitioned instances that demonstrate either typical or relatively unique behaviors and learns to classify by a boosting algorithm. It represents a well-calibrated and competitive modelling tool for handling imbalanced classification problems.

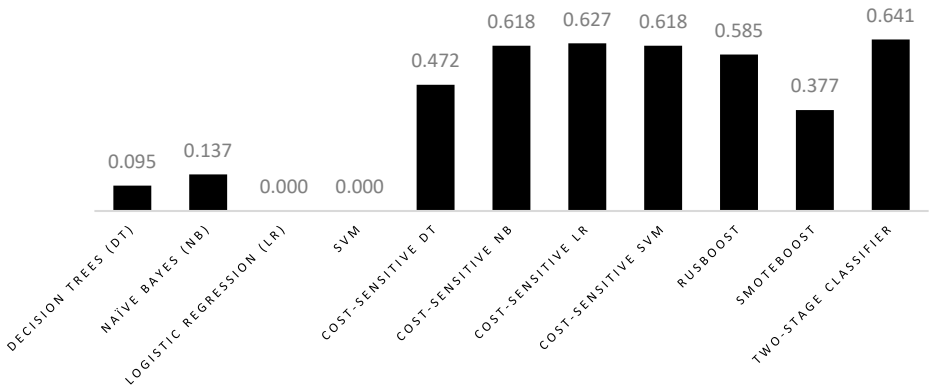

Fig. 3. The G-mean comparison between the proposed classifier and other approaches.

VI. Conclusion

We have presented a new modelling approach for predicting college freshman retention just after their first semester. Student retention especially at early times is one of the most critical problems facing by all institutions of higher education. It is also an extremely challenging classification problem to deal with due to its highly imbalanced class distributions and lack of relevant student information that can be used to differentiate and predict instances of different classes. The proposed approach applies an innovative data sampling procedure and the standard boosting algorithm to identify the freshmen who are potentially at risk of dropping out from the school. We have applied the approach to a data set that contains student information typically collected through admission processes by most universities and compared it with several widely used classification algorithms. Experimental results have indicated that the proposed approach represents a competitive and well-balanced method to combat class imbalanced problems.

Conflict of Interest

The author declares no conflict of interest.

References

[1] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and further visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3-17, 2010.

[2] F. Chacon, D. Spicer and A. Valbuena, "Analytics I Support of Student Retention and Success," *Research Bulletin*, vol. 3, EDUCAUSE Center for Applied Research, 2012.

[3] C. Yu *et al.*, "A data mining approach for identifying predictors of student retention from sophomore to junior year," *Journal of Data Science*, vol. 8, pp. 307-325, 2010.

[4] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, pp. 89-125, 1975.

[5] J. Bean *et al.*, "The psychology underlying successful retention practices," *Journal of College Student Retention*, vol. 3, no. 1, pp. 73-89, 2001.

[6] L. S. Hagedorn, "How to define retention: A new look at an old problem," in *College Student Retention: Formula for Student Success*, A. Seidman Ed., Praeger Publishers, 2005.

[7] M. S. DeBerard, G. I. Spielmans, and D .L. Julka, "Predictors of academic achievement and retention among college freshmen: A longitudinal study," *College Student Journal*, vol. 38, no. 1, pp. 66-80, 2004.

[8] W. Pan, S. Gao, C. Alikonis, and H. Bai, "Do intervention program assist students to succeed in college? A multilevel longitudinal study," *College Student Journal*, vol. 42, pp. 90-98, 2008.

[9] E. P. Jiang, "Analyzing and predicting student academic achievement using data mining techniques," *Encyclopedia of Information Science and Technology*, pp. 2453-2461, 2015.

[10] B. L. Garden, J. E. Dyers, and B. O. King, "Factors associated with the academic performance and retention of college agriculture students," *North American Colleges and Teachers of Agriculture Journa*l, vol. 45, no. 1, pp. 21-27, 2002.

[11] K. Noble, N. T. Flynn, J. D. Lee and D. Hilton, "Predicting successful college experiences: evidence from a first year retention program," *Journal of College Students Retention: Research. Theory & Practice*, vol. 9, no. 1, pp. 39-60, 2007.

[12] S. Herzog, "Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression," *New Directions for Institutional Research*, no. 131, pp. 17-33, 2006.

[13] M. Galar *et al.*, "A review of ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetic*s, vol. 42, no. 4, pp. 185-197, 2012.

[14] V. Lopez *et al.*, "An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113-141, 2013.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[16] B. Du *et al.*, "Catch me if you can: detecting pickpocket suspects from large-scale transit records," in *Proc. ACM Conference on Knowledge Discovery and Data Mining*, 2016.

[17] B. Scholkopf *et al*., "Estimating the support of a high-dimensional distributions," *Neural Computation*, vol. 13, no. 7, pp. 1443-1471, 2001.
[18] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th International Conference in Machine Learning*, 1996, pp. 148-156.
[19] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1973.
[20] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proc. 11th Conference on Uncertainty in Artificial Intelligenc*e, 1995, pp. 338-345.
[21] J. S. Cramer, "The origins of logistic regression," Technical Report, Tinbergen Institute, pp. 167-178, 2002.
[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
[23] J. Platt, "Fast training of support vector machines using sequential minimal optimization," *Advances in Kernel Methods – Support Vector Learning*, 1998.
[24] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 659-665, 2002.
[25] C. Steiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: a hybrid approach to alleviating class imbalanced," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 1, pp. 463-484, 2010.
[26] N. V. Chawla, A. Laxarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. 2003 Principles of Knowledge Discovery in Databases*, 2003, pp. 107-119.
[27] R. Reason, "Student variables that predict retention: recent research and new developments," *NASPA Journal*, vol. 46, no. 3, pp. 482-501, 2009.

Eric P. Jiang is a professor of computer science at University of San Diego, USA. In 1998, he received his Ph.D. degree in computer science from the University of Tennessee at Knoxville, USA. His research interests include information retrieval and management, data analytics, machine learning, parallel and distributed computing.