Expert Systems with Applications 41 (2014) 321–330

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition

Dech Thammasiri [a], Dursun Delen [b,*], Phayung Meesad [c], Nihat Kasap [d]

[a] Faculty of Information Technology, King Mongkut's University of Technology North Bangkok Bangsue, Bangkok 10800, Thailand
[b] Spears School of Business, Department of Management Science and Information Systems, Oklahoma State University, Tulsa, OK 74106, USA
[c] Faculty of Information Technology, King Mongkut's University of Technology North Bangkok Bangsue, Bangkok 10800, Thailand
[d] School of Management, Sabanci University, Istanbul 34956, Turkey

ARTICLE INFO

*Keywords:*
Student retention
Attrition
Prediction
Imbalanced class distribution
SMOTE
Sampling
Sensitivity analysis

ABSTRACT

Predicting student attrition is an intriguing yet challenging problem for any academic institution. Class-imbalanced data is a common in the field of student retention, mainly because a lot of students register but fewer students drop out. Classification techniques for imbalanced dataset can yield deceivingly high prediction accuracy where the overall predictive accuracy is usually driven by the majority class at the expense of having very poor performance on the crucial minority class. In this study, we compared different data balancing techniques to improve the predictive accuracy in minority class while maintaining satisfactory overall classification performance. Specifically, we tested three balancing techniques—over-sampling, under-sampling and synthetic minority over-sampling (SMOTE)—along with four popular classification methods—logistic regression, decision trees, neuron networks and support vector machines. We used a large and feature rich institutional student data (between the years 2005 and 2011) to assess the efficacy of both balancing techniques as well as prediction methods. The results indicated that the support vector machine combined with SMOTE data-balancing technique achieved the best classification performance with a 90.24% overall accuracy on the 10-fold holdout sample. All three data-balancing techniques improved the prediction accuracy for the minority class. Applying sensitivity analyses on developed models, we also identified the most important variables for accurate prediction of student attrition. Application of these models has the potential to accurately predict at-risk students and help reduce student dropout rates.



## 1. Introduction

Increasing the student retention is a long term goal of any university in the US and around the world. The negative effects of student attrition are evident to students, parents, university and the society as a whole. The positive impact of increased retention is also obvious: college graduates are more likely to have a better career and have higher standard of life. College rankings, federal funding agencies, state appropriation committees and program accreditation agencies are all interested in student retention rates. Higher the retention rate, more likely for the institution to be ranked higher, secure more federal funds, traded favorably for appropriation and have easier path to program accreditations. Because of all of these reasons, administrator in higher education administrators are feeling increasingly more pressure to design and implement strategic initiatives to increase student retention rates. Furthermore, universities with high attrition rates face the significant loss of tuition, fees, and potential alumni contributions (Scott, Spielmans, & Julka, 2004). A significant portion of student attrition happens in the first year of college, also called the *freshmen year*. According to Delen (2011), fifty-percent or more of the student attrition can be attributed to the first year in the college. Therefore, it is essential to identify vulnerable students who are prone to dropping out in their freshmen year. Identification of the at-risk freshmen students can allow institutions to better and faster progress towards achieving their retention management goals.

Many modeling methods were found to assist institutions in predicting at-risk students, planning for interventions, to better understand and address fundamental issues causing student dropouts, and ultimately to increase the student retention rates. For many years, traditional statistical methods have been used to predict students' attrition and identify factors that correlate to their academic behavior. The statistics based methods that are more frequently used were logistic regression (Lin, Imbrie, & Reid, 2009; Scott et al., 2004; Zhang, Anderson, Ohland, & Thorndyke, 2004),

* Corresponding author. Tel.: +1 (918) 594 8283; fax: +1 (918) 594 8281.
*E-mail address:* dursun.delen@okstate.edu (D. Delen).
*URL:* http://spears.okstate.edu/~delen (D. Delen).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.046

discriminant analysis (Burtner, 2005) and structural equation modeling (SEM) (Li, Swaminathan, & Tang, 2009; Lin et al., 2009). Recently, many researchers have focused on machine learning and data mining techniques to study student retention phenomenon in higher education. Alkhasawneh (2011) proposed a hybrid model where he used artificial neural networks for performance modeling and used genetic algorithms for selecting feature subset in order to better predict the at-risk students and to obtain thorough understanding of the factors that relate to first year academic success and retention of students at Virginia Commonwealth University. Delen (2010) used a large and rich freshmen student data, along with several classification methods to predict attrition, and using sensitivity analysis, explained the factors that are contributing to the prediction models in a ranked order of importance. Yu, DiGangi, Jannasch-Pennell, Lo, and Kaprolet (2007) conducted a study where they used classification trees for predicting attrition and for identifying the most crucial factors contributing to retention. Zhang and Oussena (2010) proposed data mining as an enabler to improve student retention in higher education. The goal of their research was to identify potential problems as early as possible and to follow up with best possible intervention options to enhance student retention. They built and tested several classification algorithms, including Naïve Bayes, Decision Trees and Support Vector Machines. Their results showed that Naïve Bayes archived the highest prediction accuracy while the Decision Tree with lowest one.

This brief review of the previous studies shows that data mining methods have a great potential to augment the traditional means to better manage student retention. Compared to the traditional statistical methods, they have fewer restrictions (e.g., normality, independence, collinearity, etc.) and are capable of producing better prediction accuracies. Particularly when working with large data sets that contain many predictor variables, data mining methods proven to be robust in dealing with missing data, capturing highly complex nonlinear patterns, and hence producing models with very high level of prediction accuracy. Although, there is a consensus on the use of data mining and machine learning techniques, there is hardly any consensus on which data mining technique to use for the retention prediction problem. Literature has shown superiority of different techniques over the other in variety of different institutional settings. Depending on the data, and the formulation of the problem, any data mining technique can come out to be superior to any other. This lack of consensus prompts an experimental approach to identifying and using the most appropriate data mining technique for a given prediction problem. Therefore, in this study we developed and compared four different data mining techniques.

In the retention datasets, there usually are relatively fewer instances of students who have dropped out compared to the instances of students who have persisted. This data characteristic where the number of examples of one flaw type (i.e., a class label) is much higher than the others is known as the problem of imbalanced data, or the class imbalance problem. We found that in our dataset, minority class samples constitutedonly about 21% of thecomplete dataset. According to Li and Sun (2012) if the proportion of minority class samples constitutes less than 35% of the dataset, the dataset is considered as imbalanced. Therefore, in this study we are to deal with an imbalanced class distribution problem. The class imbalance problem is not unique to student retention, it is an intrinsic characteristics of many domains including credit scoring (Brown & Mues, 2012), prediction of liquefaction potential (Yazdi, Kalantary, & Yazdi, 2012), bankruptcy prediction (Olson, Delen, & Meng, 2012) and biomedical document classification (Laza, Pavon, Reboiro-Jato, & Fedz-Riverola, 2011). It has been reported in data mining research that when learning from imbalanced data, data mining algorithms tend to produce high predictive accuracy over the majority class, but poor predictive accuracy over the minority class. Learning from imbalanced data thus becomes an important sub field in data mining research. To improve the accuracy of classification methods with imbalanced data, several methods have been previously studied. These methods could be considered as a data preprocessing that take place before applying the classification methods. The methods to balance imbalanced data sets employ some variant of under sampling and/or over sampling of the original data sets.

In this research study, we developed and tested numerous prediction models using different sampling strategies such as undersampling, over-sampling and SMOTE to handle imbalanced data. Using four different modeling techniques—logistic regression, decision tree, neural networks and support vector machines—over four different data structures—original, balanced with over-sampling, balanced with under-sampling and balanced with SMOTE—we wanted to understand the interrelationships among sampling methods, classifiers and performance measures to predict student retention data. In order to minimize the sampling bias in splitting the data between training and testing for each model building exercise, we utilized 10-fold cross validation. Overall, we executed a $4 \times 4 \times 10$ experimental design that resulted in 160 unique classification models. The rest of the paper is organized as follows: Section 2 provides a condensed literature review on student retention and the class imbalance problem. Section 3 describes the freshmen student dataset, and provides a brief review of the classification models, imbalance data techniques and evaluation metrics used for our study. Section 4 presents and discusses the empirical results. Section 5, the final section, concludes the paper with the listing of the contributions and limitations of this study.

## 2. Literature review

In this section, we first review the student retention problem from theoretical perspective—concept and theoretical models of student retention—and then review it from analytic perspective where machine learning and data mining techniques are used for classification of student attrition. In the second part of the section, we reviewed the literature on the methods used for handling class imbalance problem.

### 2.1. Student retention

There are two types of outcomes in student retention: *typical stayer* is a student enrolled each semester until graduation and graduates in due course plan; *a dropout*, or leaver, is a student who enters university but leaves prematurely or drop out before graduation and never returns to study again. High rates of student attrition have been reported in the reality of college readiness 2012 (see act.org).

Over the last several decades, researchers have developed the most comprehensive models (theoretical as well as analytic) to address higher education student retention problem. Earlier studies dealt with understanding the reasons behind student attrition by developing theoretical models. Undoubtedly the most famous researcher in this area is Tinto (1987). His student engagement model has served as the foundation for hundreds of other theoretical studies. Later, in addition to understanding the underlying reasons, the researchers have been interested in identifying at-risk students as early as possible so that they can prevent the likelihood of dropping out. Early identification of the students with higher risk of dropping out provides the means for the administrators to instigate intervention programs, provide assistance for those students in need. In earlier analytical approaches, traditional statistical methods such as logistic regression, discriminant analysis and