European Journal of Operational Research 281 (2020) 575–587

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition

Dursun Delen[a,*], Kazim Topuz[b], Enes Eryarsoy[c]

[a] Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, Stillwater, Oklahoma, USA
[b] School of Finance, Operations Management, and International Business, Collins College of Business, The University of Tulsa, Tulsa, Oklahoma, USA
[c] Department of Management Information Systems, School of Management and Administrative Sciences, Istanbul Sehir University, Turkey

ARTICLE INFO

Article history:
Received 21 February 2018
Accepted 7 March 2019
Available online 29 March 2019

Keywords:
Student retention
Prediction
Elastic net
Bayesian Belief Network (BBN)
Imbalance data

ABSTRACT

Student attrition – the departure from an institution of higher learning prior to the achievement of a degree or earning due educational credentials – is an administratively important, scientifically interesting and yet practically challenging problem for decision makers and researchers. This study aims to find the prominent variables and their conditional dependencies/interrelations that affect student attrition in college settings. Specifically, using a large and feature-rich dataset, proposed methodology successfully captures the probabilistic interactions between attrition (the dependent variable) and related factors (the independent variables) to reveal the underlying, potentially complex/non-linear relationships. The proposed methodology successfully predicts the individual students' attrition risk through a Bayesian Belief Network-driven probabilistic model. The findings suggest that the proposed probabilistic graphical/network method is capable of predicting student attrition with 84% in AUC – Area Under the Receiver Operating Characteristics Curve. Using a 2-by-2 investigational design framework, this body of research also compares the impact and contribution of data balancing and feature selection to the resultant prediction models. The results show that (1) the imbalanced dataset produces similar predictive results in detecting the at-risk students, and (2) the feature selection, which is the process of identifying and eliminating unnecessary/unimportant predictors, results in simpler, more understandable, interpretable, and actionable results without compromising on the accuracy of the prediction task.

1. Introduction

Predicting attrition (i.e., early identification of undesirable departures) has always been an intriguing and challenging problem for researchers and decision makers (i.e., practitioners, administrators or business managers). Attrition analysis—as a tool to better understand and manage retention—is an important subject in a variety of domains and based on the specific domain, it may be named differently. For instance, in marketing, it is often called "customer churn analysis" referring to timely identification of the at-risk customers (i.e., the ones about to leave you—your products and/or services—for your competitors'). The common marketing philosophy that states "acquiring a new customer is ten times as costly as retaining current customers" is a good testament to the importance of predicting and properly managing customer churn. Attrition is also an important concept in human resource management, specifically in employee retention, where accurate prediction and management of at-risk employees would save time and money to the organization in maintaining a productive and capable workforce. In this study, the focus is on student attrition, where better understanding and accurate prediction of attrition can lead to effective management of retention and graduation of students in higher education.

Student attrition is the precursory dropout of students in institutions of higher learning prior to achieving any recognized degrees or credentials (Johnson, 2012). Decreasing the attrition rate (i.e., increasing the retentions rate) provides an institution a better chance for procuring a higher status in college evaluations/rankings, potentially leading to increasing funding opportunities, recruiting better students, and having a less complicated path to program/degree accreditations. In addition to these explanations, financial loss, increased federal and state level attention make administrators in institutions of higher learning feel continually increasing levels of pressure to create and execute strategic initiatives to decrease student attrition (Golde, 2005). Nowadays, a vast majority of higher education institutions have "student success centers" and related programs and services

* Corresponding author.
E-mail addresses: dursun.delen@okstate.edu (D. Delen), kat0141@utulsa.edu (K. Topuz), eneseryarsoy@sehir.edu.tr (E. Eryarsoy).

https://doi.org/10.1016/j.ejor.2019.03.037
0377-2217/

to help retain more of their students. However, achieving low attrition rates has clearly been a difficult hurdle to overcome for many schools, and according to the American Institutes for Research's national averages, only a little over thirty percent of students admitted to the US colleges earn a college degree in higher education (Johnson, 2012). Over fifty percent of the dropouts occur within the first/freshmen-year (Delen, 2010).

Several researchers have investigated the factors affecting the student attrition and proposed theories based on qualitative or survey-based research studies where the goal was to identify and understand the underlying complex social phenomenon (Berger & Milem, 1999; Forsman, Van den Bogaard, Linder, & Fraser, 2015; Tinto, 1997). Although the research is rather disbursed and widely inconclusive, chronologically, we can group the proposed factors affecting student attrition into two main groups/categories: (1) student-related factors/characteristics before the college enrollment, and (2) social/academic experiences of the student during the early times at the college. According to the extant literature, pre-college factors such as sociodemographic characteristics, high school and prep school rankings, grade point averages (GPAs), standardized test scores (Ishitani, 2016; Pike, Hansen, & Childress, 2014); and extracurricular factors such as student social integration, financial support, and first-semester academic success seem to have direct influence in explaining student attrition (Chen & DesJardins, 2010; Tinto, 2012b). Early studies have utilized traditional statistical methods to assist institutions in understanding the reasoning behind attrition, and students at risk of dropping out of college (Berger & Milem, 1999; Forsman et al., 2015). Many have developed statistical models to discover the causality of factors leading to student dropouts (Burtner, 2005; Lee & Choi, 2013; Zhang, Anderson, Ohland, & Thorndyke, 2004). In addition to the traditional body of literature produced on the topic, in recent years, there has been an increased emphasis in analytics-based studies that utilized machine learning and data mining techniques (Nandeshwar, Menzies, & Nelson, 2011; Oztekin et al., 2016; Yukselturk, Ozekes, & Türel, 2014). Some of them (including this study) have focused on students dropped at the first year of college (i.e., the freshmen year), because of the fact that most of the college dropouts tend to occur during the freshman year (Delen, 2011; Thammasiri, Delen, Meesad, & Kasap, 2014).

The data used in retention-related studies is usually imperfect–characterized by missing values, imbalanced/skewed distributions, and highly correlated multivariate attributes. Only a few of the previous studies have considered handling imbalanced data problem, which is quite common to the field of student retention, mainly because out of all that registers, only a small percent of the students drop out, naturally creating a misbalanced representation between the numbers of dropout and non-dropout record counts in the dataset. While dealing with multivariate, imperfect datasets, a best-practice driven methodology is needed for variable selection, data transformation. Furthermore, the extent of the existing research has shown that none of the previous studies considered probabilistic models to extract the likelihood dependencies among the attrition factors, and the combined effect of the factors on the response/predicted variable/factor, student attrition.

This study analyzes the important factors and the conditional interrelations among them, collectively explaining the reasoning behind student attrition in college settings, by using a large and feature-rich institutional dataset. Specifically, this study proposes a multi-step methodology to predict an individual students' attrition risk likelihood by holistically considering the probabilistic relationships amongst potential attrition factors (i.e., independent variables). In this study, to address the data disproportion issue, we employed synthetic minority over-sampling (SMOTE) technique, which is a data balancing algorithm that assists in increasing the prognostic accurateness in minority class while also ensuring maintenance of the inclusive designation performance/accuracy. The proposed methodology also includes an elastic net (EN) algorithm to identify the most relevant variables (and eliminate the redundant/noisy variables) from the data set for development of the most efficient prediction model (Zou & Hastie, 2005). After characterizing and pruning/simplifying the scope of the research, the probabilistic relations between carefully chosen factors and attrition are designed using a data-driven, tree-augmented naïve Bayesian network (TAN) learning algorithm. Equally importantly, in this study, we also investigate how different categories of information contribute to the prediction of the student attrition by designing and implementing a sensitivity analysis procedure. After employing the 10-fold cross-validation method, the findings show the positive impact of data balancing and variable selection in overall prediction accuracy. The proposed model in this study contributes to the theory and practice of business analytics, application of which creates value to institutions in higher education. This study also contributes to the data quality related discussions in business analytics by highlighting the importance of properly handling the data disproportion issue in predictive modeling.

This manuscript is arranged in the following manner: Section 2 provides a review of the pertinent studies on student retention—with a specific emphasis on predicting and explaining the first-year/freshman student attrition. Section 3 describes the proposed probabilistic hybrid decision analytic framework—explaining the steps/tasks constituting the proposed methodology, starting with the preprocessing of the data, continuing with the designing, developing and validating the models, and ending with analyzing the models with sensitivity analysis. Section 4 shows the experimental results and deliberates on the findings. Finally, Section 5 concludes the study with a precis of the study's contributions/significance and potential future research directions.

2. Literature review

In Section 2, we begin by reviewing existing pertinent literature about attrition factors and theoretical perspective along with the survey-based methods. Then, the extant literature on traditional statistical and machine learning based prediction/classification models of student attrition are summarized. In this section, we also include literature about successful applications of Bayesian networks (BNs) in various domains and the ones about the importance of handling the class imbalance problem.

The earlier studies primarily focused on understanding the root causes behind student attritions, and the later studies mostly focused on identifying/predicting the students at risk of dropout. Most of the earlier studies were survey-driven and mostly appeared in education journals. As one of the seminal works in this field, Tinto's (1997) student engagement theory suggested that student's integration into the academic institution is a crucial factor for persistence. Tinto's student engagement model (1997) inspired others to carry out similar research studies in this area (Berger & Milem, 1999; Forsman et al., 2015). They aimed at developing statistical models to discover the factors affecting student dropouts. These factors can be summarized as pre-college factors include (a) demographics—such as age, gender, race, marital, and socioeconomic status, (b) academic—such as high school GPA, standardized achievement test (ACT or SAT) results, and credit transfer status (Felten et al., 2016; Pike et al., 2014; Tinto, 2012b). Characteristics during college include (a) academic engagement—such as fall GPA, major declared number of hour registered in fall, number of hours earned in fall, and, persistence; (b) financial support—such as student loans, grant, tuition, and scholarship; (c) institution related factors—such as, admission status, enrolled college, and degree type (Chen & DesJardins, 2010; Ishitani, 2016; Tinto, 2012a). Some of these studies utilized *p*-value based criteria and

found that some factors are statistically insignificant/negligible. For instance, the influence of major declaration or student programming in the first year was found negligible in a liberal art college setting (Howard, 2015). Even though most of the foundation build upon the survey-based studies, they seem to have two major drawbacks: lack of generalization and high cost involving in performing large-scale surveys. Furthermore, some of the conventional statistical methods were also used to forecast student attrition and detect influences that relate to their academic performance. The most commonly used models were on structural equation modeling (Lee & Choi, 2013), logistic regression (Zhang et al., 2004), and discriminant analysis (Burtner, 2005). Generally speaking, these studies tend to have low accuracy results when compared to the advanced machine learning models since they fall short of uncovering potentially complex/non-linear patterns that could affect the outcomes (Heredia, Amaya, & Barrientos, 2015).

Research studies that rely on data mining and machine learning methods are capable of handling variable-related assumptions and other restrictions (e.g., missing values, dependence, correlation, and normality), thereby capturing and representing nonlinear relationships and thereby producing better prediction accuracies. Recent studies have shown that a wide variety of data mining and machine learning techniques are adopted and successfully applied to a variety of prediction problems. Although machine learning techniques have produced superior results in many domains, student retention studies have shown somewhat mixed result—logistic regression, a statistical classification method have shown significant competence, leading to no consensus on which method is the best, and should be used, for predicting the student attrition (Delen, 2011; Heredia et al., 2015; Oztekin , 2016; Pittman, 2008; Yukselturk et al., 2014). For instance, Pittman (2008) suggested that logistic regression has a better outcome (C-statistics) than neural networks, naïve Bayes, and decision trees in predicting student attrition. Lauría et al. (2012) used undergraduate student data with oversampling-type balancing technique (Lauría, Moody, Jayaprakash, Jonnalagadda, & Baron, 2013). They developed logistic regression, support vector machines, and C4.5-type decision tree models with a balanced dataset. Based on the classification accuracy results, their model showed that logistic regression and support vector machines had better performance than the C4.5 decision tree model in identifying the at-risk students. Thammasiri et al. (2014) designed an experiment to compare various techniques for data balancing with the goal of increasing the efficacy of predictive accuracy while minority class while also retaining a suitable overall classification outcome. They found that SMOTE with support vector machines performed better than the various combination of data balancing techniques and data mining models such as logistic regression, neural networks, and decision trees. Aulck, Velagapudi, Blumenstock, and West (2016) used a balanced dataset of over 32,500 students for predicting student attrition and compared logistic regression, random forest, and k-nearest neighbor methods (Aulck et al., 2016). They found that logistic regression yielded the best performance based on C-statistics results. Lin (2012) compared different decision tree algorithms to predict student retention. They found that the best precision performance with the alternative decision tree (ADT) learning algorithm where precision was at the rate of 84% but the recall dropped to 12% (Lin, 2012). Nandeshwar et al. (2011) found that using data mining could reveal a rich level of detail about the specific school when predicting student attrition. They performed a comparison of various classification models (C4.5 decision tree, AdTrees, naïve Bayes) to predict attrition; and by analyzing the prediction models, to identify the top attributes affecting the student attrition (Nandeshwar et al., 2011).

From a longitudinal perspective, more contemporary research projects have detailed how data mining methods spawned a reasonably good performance in predicting and partially explaining student attrition. However, there is no study where the focus is to extract the conditional probabilistic relations among the potential predictors and hence to identify an individual student-level characterization of the attrition. Probabilistic graphical models (PGMs) play an important role in modeling complex nonlinear relations among the factors, as well as representing the reasoning under uncertainty. Bayesian Belief Networks (BBNs) have become popular among PGMs as these models are used to process formerly unrecognized, but potentially important data found in the networks subject to research. BBNs also actively make use of the philosophies from graph and probability theories. Researchers have developed BBN models in various domains such as prediction of failures in the rail industry caused by weather-related issues (Wang, Xu, Tang, Yuan, & Wang, 2017), software project risk analysis (Hu, Zhang, Ngai, Cai, & Liu, 2013), operational risk assessment (Barua, Gao, Pasman, & Mannan, 2016), prediction of food fraud type (Bouzembrak & Marvin, 2016), financial fraud detection (Ngai, Hu, Wong, Chen, & Sun, 2011), and the detection of diseases (Meyfroidt, Güiza, Ramon, & Bruynooghe, 2009). BNNs have also been used in higher-education for predicting student performance (Wang & Beck, 2013 and Madarshahian et al., 2017). However, to the best of our knowledge, BBN-based student attrition models that capture the conditional dependencies among predictors is new to the extant literature and hence warrants a thorough investigation.

Learning (inducing) the BBN structure from a given data set is proven to be a challenging task. Locating an ideal framework over an extensive group of data defined as an NP-hard problem because of the acyclic graph restrictions (Nielsen & Jensen, 2009). Rough answers utilizing heuristics are computationally effective, but less than ideal. The datasets used can often have excess noise or immaterial variables that create unneeded complications and misdirection for the method. Consequently, prior to developing a BBN model, an appropriate variable selection process needs to be used to remove the inconsequential variables while still acquiring the most pertinent ones (Shih, Kim, Chen, Rosenberger, & Pilla, 2014). For obtaining the most optimal batch of variables for BBN, the use of an elastic net (EN) is shown to be a good approach (Topuz, Uner, Oztekin, & Yildirim, 2017). Essentially, EN is a selection technique which utilizes a convex grouping of Ridge and Lasso regression models (Zou & Hastie, 2005). These two regularization models utilize L1 (Manhattan distance) and L2 (Euclidean distance) norms, which curtail the squared residuals, to lessen the scale of the regression model's coefficients (Tibshirani, 1996).

3. A probabilistic data analytics methodology

This study proposes a five-step probabilistic attrition risk modeling (PARM) framework for recognizing the significant variables affecting the first-year student's attrition propensity by uncovering the hidden probabilistic relations among all factors (input and output). The steps in PARM include (0) constructing dataset from various disjoint organizational data repositories, (1) preprocessing, (2) employing EN regularization technique to select the most important risk factors from the carefully preprocessed dataset, (3) building probabilistic influence diagrams and constructing the BBN for complex probabilistic interactions using the identified/important risk factors, (4) conducting sensitivity analysis and what-if analysis on risk factors using the BBN model. A graphical portrayal of the proposed methodology is shown in Fig. 1.

Step 0 denotes the time demanding process of identifying and obtaining "all" of the relevant freshmen-student-level information from disparate data sources (e.g., registrar, student aid, residential life, etc.) that exist within the university. Not to mention the intricate activities involved in requesting and obtaining the needed permissions from the administrators to have access to these data

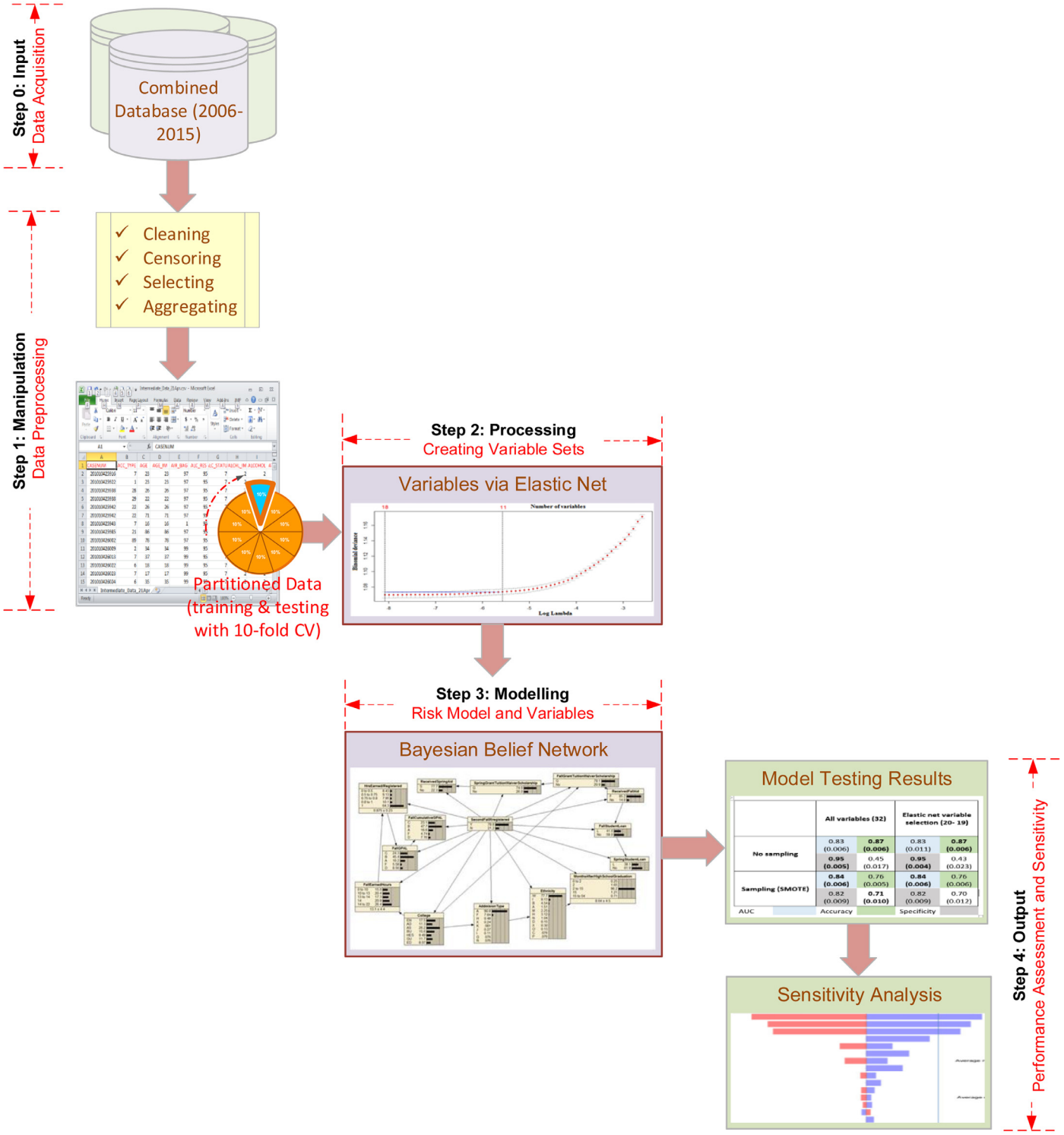

Fig. 1. A graphical depiction of the proposed research methodology.

sources. In Step 1, the combined and consolidated institutional data is preprocessed by following the "best practices" prescribed in the earlier steps of the popular data mining process CRISP-DM—Cross-Industry Standard Process for Data Mining. In Step 2, the subset of the most important predictors is identified using an elastic net-based variable selection methodology. In Step 3, the probabilistic graphical model—tree-augmented naive Bayes (TAN)—to structure the BBN inference diagram is developed. In Step 4, a sensitivity analysis is designed and executed to discover and show the level of influence portrayed by the risk factors.

In this study, an experimental design was designed and performed to illustrate the importance of variable selection and data sampling/balancing. In this 2-by-2 experimental design, four different experiments were analyzed and compared within the BBN structure: (1) all predictors with no data-balancing, (2) all predictors with data-balancing, (3) selected predictors with no data-balancing, and (4) selected predictors with data-balancing. This study used the heuristic data balancing technique, SMOTE, which was shown to have the best performance among others in the same domain (Thammasiri et al., 2014).

Moreover, to have low bias and low deviation in the performance indicators, the *k*-fold cross-validation technique was employed. In *k*-fold cross-validation, the data set is divided and placed into *k* subgroups of similar value. Every iteration utilizes *k*–1 folds as training data to create the predictive framework, and the other fold is employed as the testing data to measure the performance of the developed model. The model development and the testing process repeated *k* respectively every time, using a different fold as the test data. The cross-validation's average performance (AP) results can then be formulated as in Eq. (1):

$$\mathrm{AP} = 1/k \sum_{i=1}^{k} P_i \tag{1}$$

in which, *k* represents the number of disjointed folds/subsets, and P_i represents the tested model's performance on a given fold, *i* (Olson & Delen, 2008).

In this study, we set the value of *k* to 10. The rationale behind the choice of $k = 10$ can be explained in the following way: (a) 10-folds produce an equilibrium between the objectivity of the performance and the time/effort needed to compute the outcome of each fold, and (b) utilizing 10-folds cross-validation typically provides a balanced levels of model-comparison outcomes between bias and variance (Donate, Cortez, Sanchez, & De Miguel, 2013). An exemplary collection of the previous studies which utilized the same value (i.e., $k = 10$) include Dag, Topuz, Oztekin, Bulur, and Megahed (2016), Sevim, Oztekin, Bali, Gumus, and Guresen (2014), and Oztekin, Delen, and Kong (2009).

3.1. Data

To properly carry out the proposed experimental design, an official freshmen student dataset obtained from a public university was used. The dataset was constructed by identifying and merging data items/variables (freshmen student-related information) from several disparate databases maintained and used by the institution. The resulted combined dataset consisted of 36,461 records (3000–4000 record/student per year) and 33 factors/predictors encompassing for ten years between 2006 and 2015. Of the 36,461 records (i.e., students), roughly 20% of them did not return for the second year (the second fall semester), hence labeled as dropout freshmen students. Although the ones determined for sure were excluded from the final dataset, some of these students may have transferred to other universities without explicitly notifying the university, and hence assumed herein as dropouts. Since the proportion of these anomaly cases is very small, their impact on the analytics modeling results is deemed to be negligible. This rather large and feature-rich data included variables related to students' demographic information, pre-college characteristics, college performance factors, scholarship, and financial support specifics. A summary of the data fields for the freshmen students used in this study is given in Table 1.

During the data preprocessing phase, records with missing values for Fall Registered Hours and Fall Earned Hours (3 cases) and Spring Registered Hours (28 cases) were removed from the dataset. Since our analysis includes students that start in the fall semester, the spring semester enrollees were also removed from the dataset (738 cases). Additionally, international students' data records were removed (671 cases) for two reasons: (1) almost all of them had missing values for pre-college academic variables (SAT/ACT score and high school GPA), and (2) they have extremely low attrition rate due to legal/immigration and financial reason (they spend a prodigious sum of money coming to the US to study) (Andrade, 2006). Thus, the study included only the US citizens, and therefore, the variables like TOEFL, Visa status, and Citizenship variables were not included in the dataset. After the preprocessing steps, the final data consisted of 35,021 records.

The data preparation process included proper representation/transformation of the existing variables as well as the creation/formation of new features/variables. Accordingly, a new variable was formed based on the relativeness of *Earned Hours* and *Registered Hours* for the first/fall semester. The student completing the registered hours successfully was given an index of one. Otherwise, the student was provided with an index based on the formula: Earned Hours / Registered Hours. Also, we chose to keep both *Residency* and *Permanent address state* variables, since they may represent slightly different information about the student residency status—while *Residency* determines the amount of tuition paid, *Permanent address state* denotes whether the permanent address of the student is the school's state, neighboring state, or any other state. Financial variables including financial aid, scholarship, loan, and work-study had many incomplete/blank values. The replacement of absent values was done through further investigation of the data sources and consultation with the domain experts. The remaining records with missing/blanks values on financial variables were assumed to belong to the students who either did not apply or did not get the specific financial support. Therefore, these student records were considered as non-receiver of financial support and the blank values were replaced with "N" (i.e., none).

3.2. Variable selection with elastic net

Prior studies have proven that utilizing the ordinary least squares (OLS) regression method for variable selection has drawbacks with a large number of the coefficients that result in a reduced bias and increased variance, resulting in poor forecast precision and consistency (Tibshirani, 1996). To overcome this drawback of OLS, Lasso and Ridge are proposed, which are regularization methods aiming at lessening the summation of squared residuals using both L1 and L2 norms, which restricts the expansion of the regression model coefficients (Tibshirani, 1996). Ridge regression lessens the coefficients and their values become relationally comparable when the variables are associated. Ridge regression uses L2 norm penalized least squares (Friedman, Hastie, & Tibshirani, 2010) as represented in Eq. (2):

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right].$$

Ridge outpaces Lasso when observations exceed variables in number, or when the variables show increased correlation; however, Ridge cannot select the variables automatically (Zou & Hastie, 2005). Lasso, on the other hand, has the capabilities of providing variable selection automatically, and reduction for the coefficient of the predictor by using L1 norm penalized least squares (Tibshirani, 1996) as in Eq. (3):

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right]$$

When two or more variables are decidedly connected, it selects one variable and disregards the others. EN is a regularization method that has the capability to overcome the shortcomings of Ridge and Lasso regression, and hence, can be used for variable selection. EN utilizes a tuning parameter that provides a convex grouping of the two aforementioned regression methods (Zou & Hastie, 2005). The EN selects variables similar to the way Lasso does, however, while doing so it also eliminates any corruptions or unwanted outlying selections because of robust correlations (Friedman et al., 2010). It provides a sparse framework with a respectable prediction accuracy.

Suppose the data have n observations with p predictors, then $x_i = (x_{i1}, x_{i2}, \ldots, x_{ij})^T$ is the prediction set where $i = 1, 2, \ldots, n$

Table 1
Summary of the variables obtained from student records.

Variable	Description	Data Type	Descriptive Statistics*	%Missing/ Unknown
AdmissionType	Admission type	Nominal	A: 91, F:8	0.00
Age	Student age	Numeric	18.70 (0.71)	0.00
ClepHrs**	College-Level Examination Program (CLEP) hours	Numeric	0.06 (0.70)	0.00
College	College	Nominal	Art & Science: 24, Engineering: 18	0.00
Degree	Degree	Nominal	Bachelor: 26, Bachelor/Science: 24	0.00
Ethnicity	Ethnicity	Nominal	White: 76, Native American: 7	0.00
FallCumGPAL	Fall cumulative GPA	Numeric	2.90 (0.94)	0.00
FallEarnedHrs	Fall earned hours	Numeric	12.49 (0.94)	0.00
FallFederalWorkStudy	Fall federal work study	Binary	No: 97, Yes: 3	0.00
FallGPAL	Fall GPA	Numeric	2.84 (0.98)	0.00
FallGrantTuitionWaiverScholarship	Fall grant or tuition waiver or scholarship	Binary	Yes: 80, No: 20	0.00
FallRegisteredHrs	Fall hours registered	Numeric	14.25 (1.75)	0.00
FallStudentLoan	Fall student loan	Binary	No: 58, Yes: 42	0.00
FallWeightedGPA	Fall weighed GPA	Numeric	2.67 (1.10)	0.00
Gender	Gender	Binary	Female: 52, Male: 48	0.00
Hicomp	SAT score - combined	Numeric	24.40 (3.95)	6.92
Hiengl	SAT score - English	Numeric	24.19 (4.90)	6.92
Himath	SAT score - Math	Numeric	23.46 (4.51)	6.92
Hiread	SAT score - Reading	Numeric	25.25 (5.09)	6.92
Hisci	SAT score - Science	Numeric	24.04 (4.07)	6.92
Hsgpa	High school GPA	Numeric	3.52 (0.38)	0.00
MaritalStatus	Marital status	Binary	Single: 99, Marr.: 1	0.00
MonthsBetweenHSGradandStart	Months after high school	Numeric	1.52 (5.80)	0.00
PermAddressStateCat	Permanent address state	Nominal	School state: 70, Neighbor state: 25	0.05
PersistanceFall	Earned by registered	Numeric	0.88 (0.24)	0.00
RecvdFallAid	Fall financial aid received	Binary	Yes: 86, No: 14	0.00
RecvdSpringlAid	Spring financial aid received	Binary	Yes: 77, No: 23	0.00
Rescode	Residency	Binary	Resident: 71, Not resident: 29	0.00
SecondFallRegistered	Registered and attended the second fall semester	Binary	Yes: 80, No: 20	0.00
SprgRegisteredHrs	Number of spring hours registered	Numeric	14.78 (2.13)	29.10
SpringGrantTuitionWaiverScholarship	Received grant or tuition waiver or scholarship for spring semester	Binary	Yes: 75, No: 25	0.00
SpringlFederalWorkStudy	Received federal work study for spring semester	Binary	No: 96, Yes: 4	0.00
SpringStudentLoan	Received student loan for spring semester	Binary	No: 61, Yes: 39	0.00
TransferHrs	Number of transferred hours	Numeric	2.99 (6.40)	0.00

* *Descriptive Statistics—Numeric: mean (standard deviation); Binary:% frequency of each class; Nominal:% of most frequent 2 classes.*
** *CLEP offers exams that cover intro-level college course material. With a passing score on one CLEP exam, the student could earn three or more college credits at more than 2900 U.S. colleges and universities.*

Algorithm 1
The proposed Elastic Net-based variable search algorithm.

SET
$\tilde{\alpha} = 0$, ss= step size
While (
$\tilde{\alpha} \leq 1$
for $i = 1$ *to* k= *number of folds,*
$E_i(\tilde{\alpha}) = \arg\min_{\beta} f(\beta, \tilde{\alpha}, \lambda)$;
$\tilde{\alpha} = \tilde{\alpha} + ss$
)
FEATURE SET: Use features within one standard deviation of the average cross-validated error for each alpha value.

and $j = 1, 2, \ldots, p$, $y=(y_1, y_2, \ldots, y_n)^T$ is the response variable, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the coefficient set. EN can be formulated as in Eq. (4):

$$E = \arg\min_{\beta} f(\beta, \alpha, \lambda)$$
$$= \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha |\beta_j| \right] \right\}$$

The selection of the tuning parameter, alpha (α), depends on the level of multicollinearity in the data set. In this study, we designed an algorithm using the iterative search algorithm to find the best possible alpha value, as described in Algorithm 1:

In Algorithm 1, the parameters—lambda (λ) and beta (β)—values are determined by utilizing the minimum average cross-validated error (CV-E) and are calculated separately for each alpha value, where alpha is between 0 and 1, as the compliments of the search procedure. Utilizing Zou and Hastie's (2005) elastic net algorithm with the package called "glmnet" in R, we find the set of features as discussed in Section 4. The reason for employing just a single standard deviation error tolerance in variable selection is that having a lower number of variable choices overall provides a more regularized method that could theoretically result in more accurate and consistent prediction outcomes (Friedman et al., 2010). In every stage of the model, k-fold cross validation is implemented, and average cross-validated binomial deviance (CV-E) is measured. This study involved 33 variables, most of which were correlated to one another. The results of the proposed feature selection method are presented in Fig. 3 and in Section 4.

3.3. Bayesian Belief Network prediction model

This study employs BBN to capture and present the interrelations between independent variables and student attrition (the dependent variable). BBNs are powerful mathematical models that capture the probabilistic graphical dependency structure among several variables and variable groups in an intuitive and explicit way. At its core, the BBN model is a directed acyclic graph where the nodes correlate to the variables, and the conditional dependencies between groups of variables are represented by arcs (Pearl, 2009). This makes BBNs very useful, as they are able to employ their abilities assisting in uncertain conditions with their reasoning while also modeling complex nonlinear interactions, and much more (Koller & Friedman, 2009). In recent years, the use of BBN has become more prevalent in the literature, given its capability to manage prior unknown, yet potentially important data including topic areas like prediction of risk/survival (Dag et al., 2016), no-show risk prediction (Topuz et al., 2017), social network analysis for prevention of money laundering (Colladon & Remondi, 2017), and making medically sound choices for optimal treatment plans as well (Lucas, van der Gaag, & Abu-Hanna, 2004). BBN chain rule is a compact way of formulating complex probability distributions where each x_i represents a variable and Pa_{x_i} is the parents of the same (x_i) variable (Koller & Friedman, 2009):

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa_{x_i}). \quad (5)$$

There are two competing methods to construct a BBN model: (1) a manual method with the help of an expert, and (2) an analytical method by learning/inducing the construction from the data by utilizing innovative mathematical techniques (Koller & Friedman, 2009). Creating a modest-sized network manually needs an experienced and knowledgeable BBN engineer allocating a significant amount of time with one or more domain experts since the number of conditional probability estimations that need to be made/calculated for each parent nodes increases exponentially. For instance, a single node with n parents requires a conditional probability table that includes 2^{n+1} estimated parameters (Korb & Nicholson, 2010). Also, the expert judgment may lead to subjective decisions and may result in inexact/uncertain outcomes due to differences in experts' opinions.

The studies conducted previously offer numerous methods that utilize datasets for studying/inducing the model. The naïve Bayes classification is a relatively uncomplicated method that infers provisional freedom among all forecaster variables within the specified group or focus (Friedman, Geiger, & Goldszmidt, 1997). The classification follows the Bayes rule, which mandates to compute the likelihood of the class/target numeral for every provided value, followed by the greatest calculated prediction being selected for the structure going forward. The TAN method, and improved upgrade of the naïve Bayes classifiers, implements a tree-like model and infers the relationships between different predictor variables (Friedman et al., 1997). Furthermore, the TAN method employs a class variable without parents, while every predictor variable includes a parent with a maximum of one attribute. Consequently, an arc between two variables implies that there is a relationship between them (see Fig. 2). The TAN method tends to perform better than the other constraint-based structural learning algorithms such as Naive Bayes and Markov blanket (MB). More discussion and details on BBNs, naive Bayes, TAN, and MB can be found in the work of Korb and Nicholson (2010). In our case, we utilized the Bayesian Belief Network's TAN model because of its better performance compared to alternatives. Our ultimate goal with this model was to discover conditional dependencies among the predictors of student attrition. A proper depiction of the parents for a variablex_i

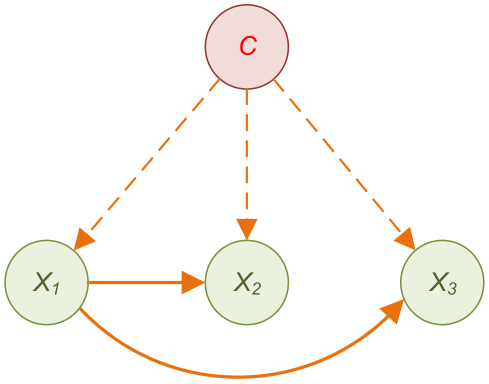

Fig. 2. Tree augmented naïve Bayes network structure.

is:

$$Pa_{x_i} = \{C, x_{\delta(i)}\} \quad (6)$$

where the tree is a function over $\delta(i) > 0$, Pa_{x_i} is the set of parents for each x_i, and C is the class variable that has no parents, namely $Pa_C = \emptyset$.

Chow and Liu (1968) describe a procedure for building this tree structure. "This procedure reduces the problem of constructing a maximum likelihood tree to a maximum weighted spanning tree in a graph" (Friedman et al., 1997).

One other important notion that plays a key role in constructing the tree is mutual information. To construct the tree, the correlation between each pair of variables in the system are measured and add edges only between those variables that are highly correlated. The measure of the correlation between two variables that forms the weight of an edge in the graph is called mutual information. The mutual information between two random xi=Fall student loan and xj= Spring student loan variables is defined as:

$$I_P(x_i; x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i) P(x_j)}, \; i \neq j \quad (6)$$

The procedure for constructing a TAN uses Chow and Liu's (1968) tree Bayesian concept. Finding a maximal weighted spanning tree in a graph is an optimization problem, where the objective is to maximize the log-likelihood of δ(i) (Chow & Liu, 1968). Then, the TAN construction steps can be as follows (Friedman et al., 1997):

Step 1: Compute conditional mutual information function for each (i, j) pairs as:

$$\mathbf{I_P(x_i : x_j | C)} = \sum_{\mathbf{x_i, x_j, C}} \mathbf{P(x_i, x_j, C) \log \frac{P(x_i, x_j | C)}{P(x_i | C) P(x_j | C)}}, \; \mathbf{i \neq j} \quad (7)$$

Where C is the class variable in our case Second fall registered. This function indicates how much information is provided when the class variable is known.

Step 2: Build a complete undirected graph, and use the conditional mutual information function to annotate the weight of an edge connecting xi to x_j.

Step 3: Build a maximum-weighted spanning tree.

Step 4: Convert the undirected graph to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.

Step 5: Construct a TAN model by adding a vertex labeled by C and adding an arc from C to each x_i.

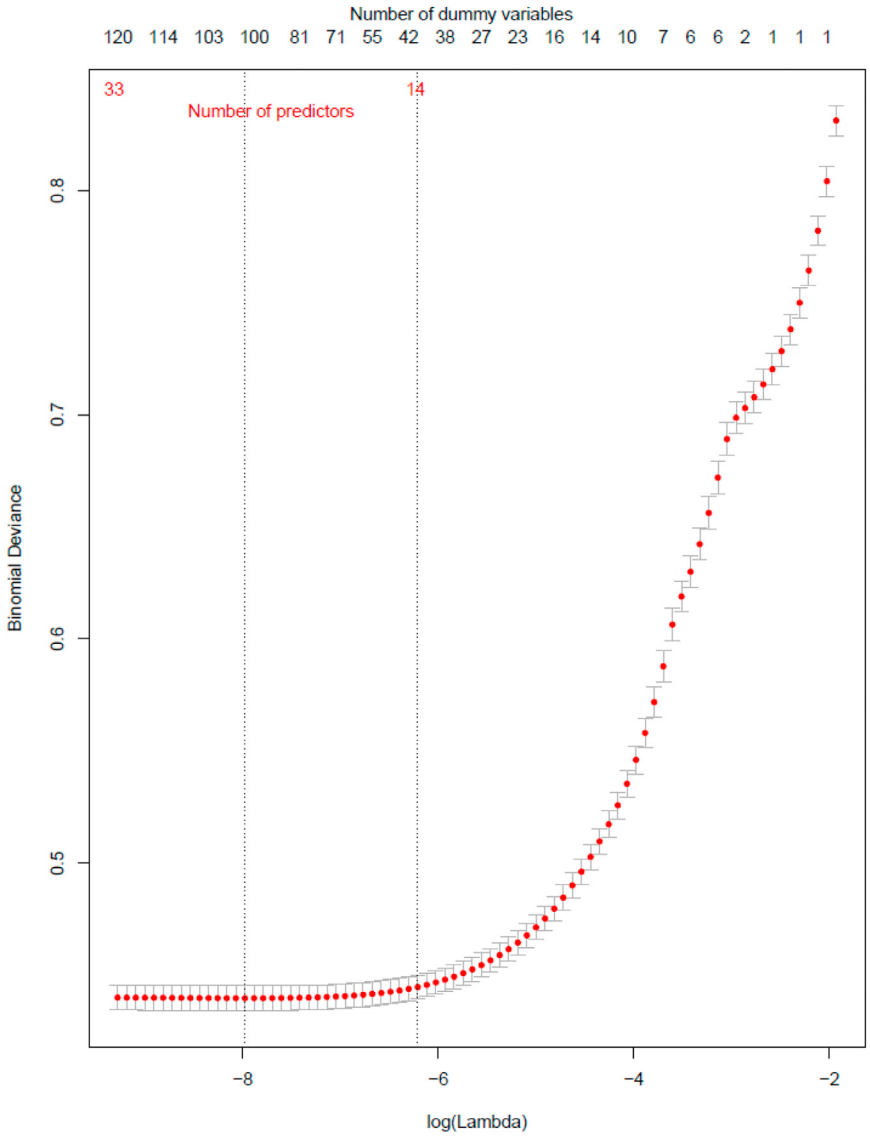

Fig. 3. Cross-validated deviance (red dotted line) with standard deviation bars (grey lines).* Dummy variables: Nominal valued variables converted to binary numeric variables using one-of-n transformation. * Binomial deviance: Exponential loss function which continuously penalizes increasingly negative margin values more heavily than they reward increasingly positive ones. For more information please see Friedman, Hastie, and Tibshirani (2001). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

After the construction of the tree, the conditional probability of each of the attributes conditioned on its parent and class label is calculated and stored. Also, the conditional probability of the root variable conditioned on class is computed and stored. Then, the posterior probability for each of the class label: $P(C \mid X_1, ..., X_n)$, is calculated as a product of the prior probability, the conditional probability of root variable and the conditional probability of all the attributes. The class label with the maximum posterior probability value is assigned to the test sample. The proof and detailed explanation regarding the TAN classification algorithm is given by Friedman et al. (1997).

3.4. Performance evaluation

Several approaches can be found in the literature regarding the "true performance" of two-class response variable models. To provide a comprehensive performance picture, three comparison categories were adapted: (a) the rate of correctly classified examples (accuracy), (b) the capability to identify the student of each group (sensitivity and specificity), and (c) the metrics for evaluating the overall performance (the area under the receiver operating characteristic curve (AUC), and the geometric mean (G-mean)). The majority of these metrics were calculated by employing the confusion matrix, which is a cross-tabulation of correctly and incorrectly predicted examples for each class (see Table 2).

A confusion matrix for a two-class (i.e., binary) classification model contains four populated cells: True positives (TP)—representing the number of accurately predicted positive cases; True Negatives (TN)—representing the number of accurately predicted negative cases; False Positives (FP)—representing the number of negative cases that are incorrectly predicted as positive; and False Negatives (FN)—representing the number of positive cases that are incorrectly predicted as negative. In essence, the diagonal cells from the upper left to lower-right contain the accurately predicted cases, while all other cell containing the incorrectly predicted cases. Below are the performance measurements used in assessing the prediction methods:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Table 2
A schema of the confusion matrix for the proposed binary classification model.

		Predicted Class Positive (student fall registered=yes)	Negative (student fall registered =no)
Actual Class	Positive (student fall registered =yes)	True positive (TP)	False negative (FN)
	Negative (student fall registered =no)	False positive (FP)	True negative (TN)

Table 3
Selected features by EN regularization.

	Features	
Admission type	Fall GPA	Received fall financial aid
College	Fall grant/ tuition waiver/ scholarship	Received spring financial aid
Ethnicity	Fall student loan	Residency*
Fall cumulative GPA	Months after high school	Spring grant/ tuition waiver/ scholarship
Fall earned hours	Earned by registered hours in fall	Spring student loan

* Residency was selected in balanced dataset only.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \text{ and Specificity} = \frac{TN}{TN+FP} \tag{9}$$

$$\text{G} - \text{mean} = \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{TN+FP}} \tag{10}$$

Additionally, AUC (area under the receiver operating characteristics [ROC] curve) acts as an inclusive performance measurement frequently utilized in studies for diagnostic testing and analyzing the performance of binary classification models (Delen, 2015). To calculate AUC, the *sensitivity* in function of the false positive rate (1-*specificity)* for different endpoints. Each point on the curve signifies a sensitive pair relating to a specific decision threshold. AUC is a measure of a given classification model's ability to discern between two groups (in the case of this study, it shows a model's discerning ability between dropouts and persistent freshmen student groups).

4. Experimental results and sensitivity analysis

4.1. Model results

It is recognized that the choice of alpha relies on the level of multicollinearity in the data set. In alpha search, the step size as 0.33 because the beginning dataset contains this specification of variables (33). The step size can be selected as a smaller value, dependent on a smaller specification of variables, and also varying alpha values can be used in a dataset when a large number of variables (greater than 100) is involved. After a process of searching and experimenting with a range of possible alpha values and comparing the corresponding results, we settled on using an alpha of 0.99, so that the model could perform like Lasso regression and, at the same time, would eliminate any odd behavior caused by the correlations (Zou & Hastie, 2005). The predictors selected by this regularization technique are shown in Table 3.

The cross-validated results of all four models on three comparison categories are listed in Table 4, where each cell contains the average and standard deviation of the respective performance measures. As the results indicate, no single dataset option (out of the four developed and compared in this study) outperformed the others in all measures.

Accuracy measure seems to be rather sensitive to the balancing but not as much to feature selection. As the results indicate, the best accuracy and specificity producing models are obtained from the imbalanced datasets. However, the sensitivity is better with the balanced datasets and both sensitivity and specificity not significantly affected by feature selection. Fig. 4 shows, within per-class accuracy measures, the prediction model produces meaningfully improved results on forecasting the "No" class with the imbalanced data than it did with balanced data (specificity of 95% vs 82%). Based on the overall metric, AUC, balanced datasets were slightly better than the imbalanced ones (0.84 vs 0.83). Feature selection helped in reducing the number of predictors by more than 50% and yet provided very similar performance outcomes (33 vs 15 or 33 vs 14) for all metrics.

Given that our goal is to predict the at-risk students, finding students that are not registered in the following fall (i.e., SecondFallRegistered = No), which is denoted with the specificity measure, is very crucial. Since the overall metric, AUC, have not improved significantly with the balanced dataset and balancing change the prior probabilities and thereby the intuitive structure of the network model, this study suggests using the network structured with the imbalanced dataset (Fig. 5). Some studies (Fallahi & Caferi, 2011; Sun, Wong, & Kamel, 2009; Thai-Nghe, Drumond, Krohn-Grimberghe, & Schmidt-Thieme, 2010; Topuz, Zengul, Dag, Almehmi, & Yildirim, 2018) have shown that BBN indeed can handle the imbalanced dataset compare to other data mining/machine learning models (e.g. logistic regression, decision trees, and neural networks). Please see Appendix A for the balanced dataset BBN network.

Fig. 5 shows the constructed BBN graphical model, exhibiting the relationships and conditional dependencies (766 total) among all predictors (14 features). Please view Appendix B for a comprehensive listing of conditional probabilities. The constructed BBN

Table 4
Ten-fold cross-validation classification performance measures for all models.

	No feature selection & imbalanced data	No feature selection & balanced data (SMOTE)	Feature selection & imbalanced data	Feature selection & balanced data (SMOTE)
Accuracy	0.87 ± 0.006	0.76 ± 0.005	0.87 ± 0.006	0.76 ± 0.006
Sensitivity	0.45 ± 0.022	0.71 ± 0.010	0.43 ± 0.023	0.70 ± 0.012
Specificity	0.95 ± 0.005	0.82 ± 0.009	0.95 ± 0.004	0.82 ± 0.009
AUC	0.83 ± 0.010	0.84 ± 0.006	0.83 ± 0.011	0.84 ± 0.006
G-mean	0.65 ± 0.017	0.76 ± 0.006	0.64 ± 0.018	0.76 ± 0.007
No. of features	33	33	14	15

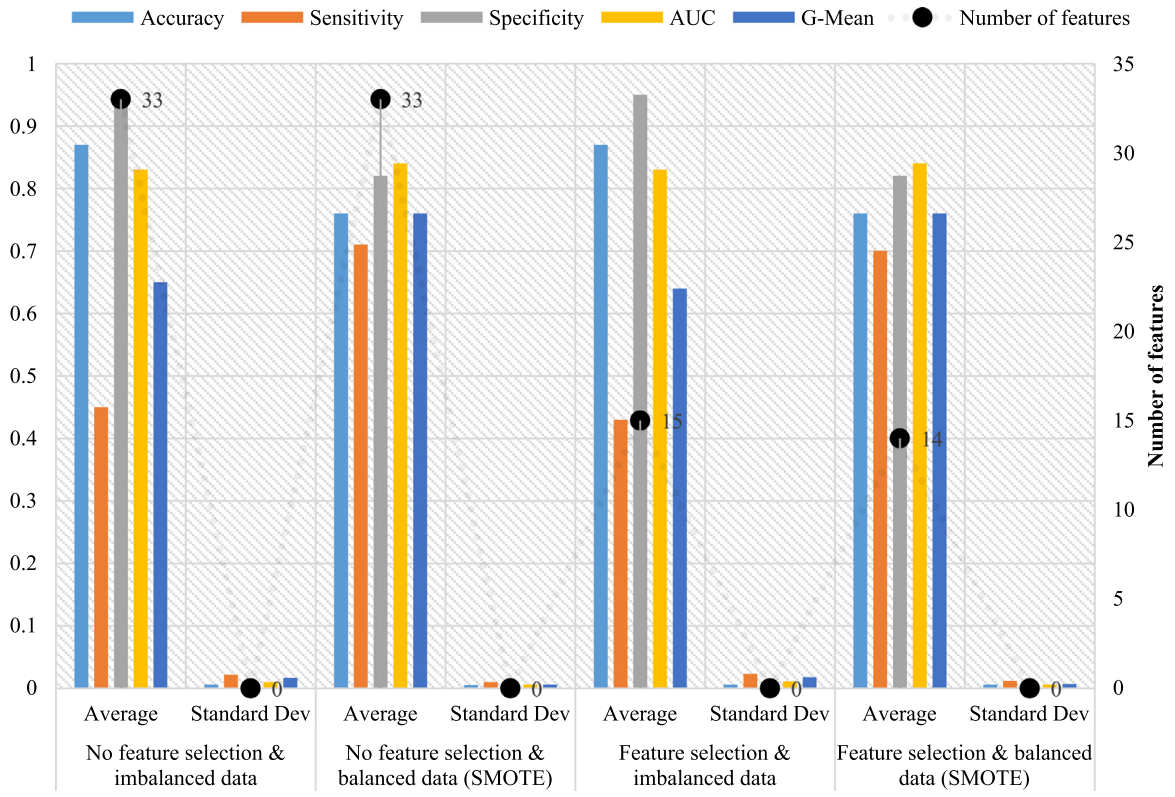

Fig. 4. Performance comparison for different datasets.

can be of great use to practitioners (i.e., administrators and managers in educational institutions) because it offers a holistic view of all relationships, and provides the means to explore detailed information using the "what-if analysis." In fact, with this network model, it is possible to calculate the student-specific risk probability of attrition, which is the posterior probability of a student who would drop out of the university, if the values of a selected/given predictor perturbed within its value domain. In this study, a *10*-fold cross validation methodology was employed; that is, we developed 10 BBN models, each time using a different

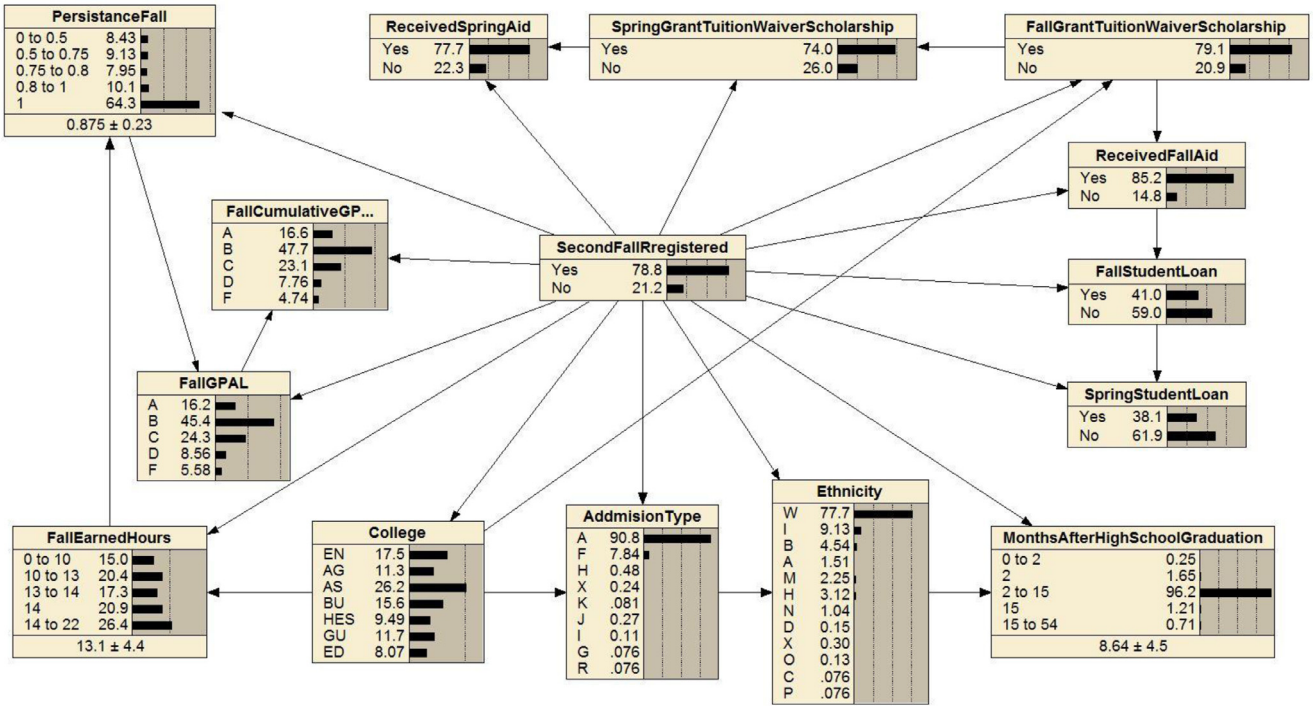

Fig. 5. Bayesian Belief Network for prediction of student attrition.

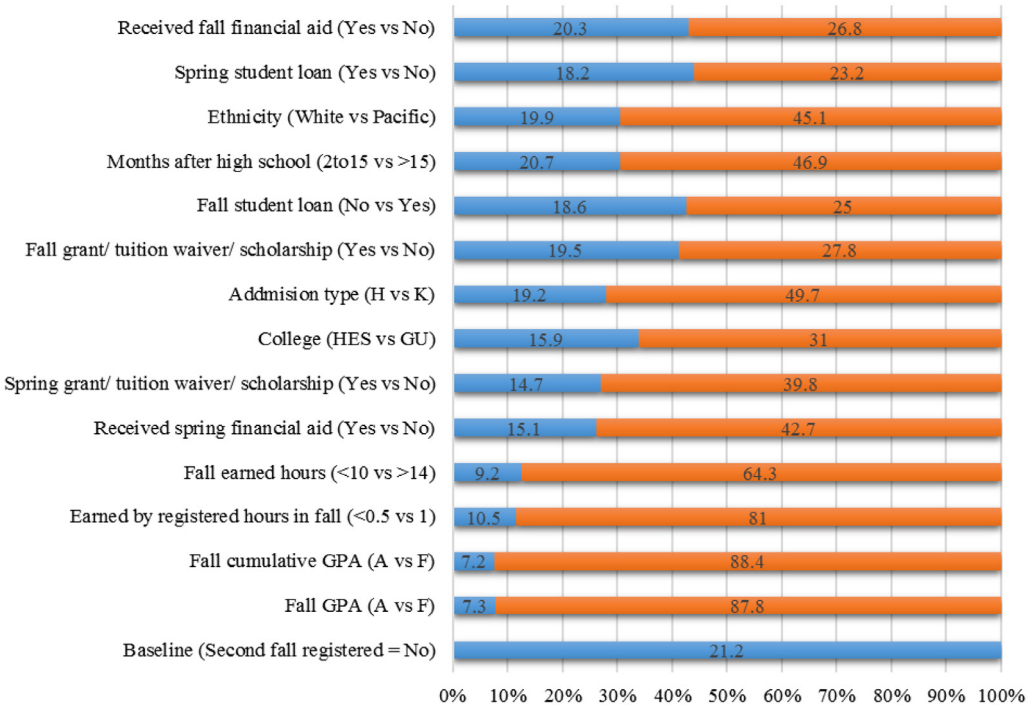

Fig. 6. Probability of student attrition for risk factors—what-if analysis on individual factors.

tenth of the complete dataset as the test set. For practical reasons, we had to pick one of the *10* folds to build the exemplary model and the corresponding graphical network. Because of its close performance to the aggregated results build on all *10* folds, we chose fold number four as the data source for the representative model. The network build using this data structure is shown in Fig. 5.

When interpreting the BBN model shown in Fig. 5, one should consider the arrows, directions of the arrows, interactions, and indirect relationships. For example, the *fall grant/tuition waiver/scholarship* category (i.e., FallGrantTuitionWaiverScholarship) and all nodes linked to FallGrantTuitionWaiverScholarship are related to student attrition (i.e., SecondFallRegistered= No). Moreover, while FallGrantTuitionWaiverScholarship interacts with college (College) and spring grant/tuition waiver/scholarship (i.e., SpringGrantTuitionWaiverScholarship) directly, it also interacts with admission type (AdmissionType) indirectly through College. According to the BBN model, another indirect connection exists between predictors *earned by registered rate* (i.e., PersistanceFall) and College. As such, if the PersistanceFall of the student is less than 0.8, then College type has an effect on student attrition. However, if the PersistanceFall of the student is 1, then the College type does not impact the student attrition in a noteworthy manner.

4.2. Sensitivity and what-if analyses

A paramount advantage of creating this type of BBN model and showing the interplay among the probabilistic factors is the ability to systematically investigate parameter uncertainty and/or sensitivity. For instance, in the fault diagnosis domain, a researcher may seek to find which nodes will provide the greatest amount of information in establishing the understanding of the nodes with the greatest probability of fault. Of course, this will change when the results are formulated, so this step may need to be reworked at each stage of the study. By showing how the beliefs and mean value of the target node could cause manipulation through an individual discovery at every other node in the network, the BBN gives a consistently useable method to update the accessible information. By definition, *belief* means posterior probability (predicated on all discoveries currently entered).

Table 5
The numerical findings of the sensitivity analysis procedure.

Node	Mutual	Percent	Variance of beliefs
Fall GPA	0.1636	21.90	0.0444
Fall cumulative GPA	0.1571	21.10	0.0426
Earned by registered hours in fall	0.1515	20.30	0.0415
Fall earned hours	0.1310	17.60	0.0353
Received spring financial aid	0.0510	6.84	0.0132
Spring grant/ tuition waiver/ scholarship	0.0479	6.42	0.0121
College	0.0075	1.00	0.0018
Admission type	0.0058	0.77	0.0015
Fall grant/ tuition waiver/ scholarship	0.0046	0.62	0.0011
Fall student loan	0.0043	0.57	0.0010
Months after high school	0.0040	0.53	0.0011
Ethnicity	0.0037	0.49	0.0009
Spring student loan	0.0026	0.34	0.0006
Received fall financial aid	0.0022	0.29	0.0005

The sensitivity of the findings table, Table 5, displays which variable gives the most accurate data about the existence of the target class (student readmission – Second fall registered). The computation to find this information can be done through the entropy function like so:

$I = H(Q) - H(Q|F)$, where H(Q) is the entropy of Q before findings, and H(Q|F) is the entropy of Q after findings. Additional information regarding the sensitivity of the results is perusable in other sources documenting the same topic (Pearl, 2014). According to the results, the most important/influential features were Fall GPA, fall cumulative GPA, and earned by registered hours in fall, where their individual relative contribution is more than 20% in identifying the students at risk of dropout. Variables including received fall financial aid and spring student loan seem to have a lower level of influence/contribution in determining the attrition. The reason behind this could be the potential of having the same in-

formation by other financial variables such as grant/tuition waiver/scholarship in fall and spring.

The next section extends the baseline case by exploring some possible "what if scenarios" in the sequence of events associated with the student's attrition. For the "what if scenario," we explored some possible impacts of these propositions to develop a conditional narrative. Fig. 6 summarizes the most positive and most negative levels within each predictor with their posterior probabilities. For instance, getting an "A" for the Fall GPA decreases the posterior probability of student attrition to 7.3%, or conversely, getting an "F" increases the probability of attrition to 87.8%, where the baseline is 21.2%.

5. Conclusion and future research directions

The current investigation sought to define the risk probability in student attrition, explore the factors that could affect the retention/attrition of a student, and reveals the potential interactions among the predictors and their effect on attrition risk. To achieve these goals, a multi-phased process was followed including conventional statistical modeling and Bayesian Belief Network.

Our results indicate that probabilistic graphical models, BBNs, can predict student attrition with approximately 83% AUC when large enough dataset with proper variables is utilized. We found that the imbalanced dataset (compared to balanced) produced better predictive outcomes (Specificity and Accuracy) for recognizing the students who are most likely to become victims of attrition precursor to their sophomore year. Moreover, with feature selection, the unnecessary predictors (more than half of them) can be eliminated and yet the final model can provide similar results, which can help administrators to focus on only the important factors.

The noteworthy strengths of this research are two-fold: first, the proposed probabilistic methodology, which combines EN variable selection with the probabilistic graphical network to allow not only prediction but also deeper understanding of the phenomenon as well as providing a model for sensitivity analysis between the predictive factors and the target variable. Second, providing a decision aid tool for administrators and domain experts, augmenting their knowledge to create better outcomes by identifying the probabilistic dependencies between the important factors and thereby revealing the hidden information about the underlying non-linear relationships among the predictors. Furthermore, our study differs from other student attrition studies that focus solely on prediction since our methodology aims to uncover the conditional relations among the predictors of attrition, to determine the conditional dependencies among the predictors, as well as to construct student specific attrition probabilities.

A positive outcome in any analytical project strongly requires rich depth (quantity and dimensionality) in the data that is representative of the underlying situations being analyzed. Although this study employed a large amount of sample data (10 years of freshman student records) with an expansive group of features (over 30 total features), the enhanced quantity and quality of relevant variables has the potential to expand the results of analytical prediction. Of the variables in the study, the ones that have shown to have the greatest potential of improving predictive ability are: the level of social interaction experienced by the student, such as membership in a fraternity or club; the financial and educational backgrounds of the student's parents and significant other; and the student's hopes and prospects before beginning his pursuit of education. Second, the findings of our study pertaining to the predictors and their conditional dependencies should be interpreted by education experts and examined in a managerial context for their practicality. Input from domain experts (administrators in educational settings) cannot and should not be ignored in any student attrition study as they can only strengthen the selection of the inputs and interpretations of the outcomes. Third, despite the methodological efforts during the data cleaning and transformation stage, it is known that the results of any retrospective analytics study built on secondary data sources would only be as good as the veracity of the data source. Although this fact may apply to this study, the data set used herein is from a well-established, reliable, and carefully designed data source, identified, collected, and verified by the subject experts. Last, while this study performed and compared BBNs structured by TAN, it should be pointed out that other BBN algorithms could have also been utilized Chickering (2002), Käser, Klingler, Schwing, and Gross (2017). To address this limitation, we compared TAN with other BBN algorithms such as Naive Bayes and Markov Blanket. In those pilot experiments, we found that TAN performed better than other constraint-based learning algorithms in all scenarios. Taken together, the findings of our study can offer useful tools for higher education administrators that can be incorporated into their decision-making process, which can potentially improve the overall retention rate and thereby enhance students' educational outcomes and the ranking/reputation of the institution.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2019.03.037.

References

Andrade, M. S. (2006). International students in English-speaking universities: Adjustment factors. *Journal of Research in International Education, 5*(2), 131–154.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. In *Proceedings of the international conference on machine learning's data for good workshop 2016*.

Barua, S., Gao, X., Pasman, H., & Mannan, M. S. (2016). Bayesian network based dynamic operational risk assessment. *Journal of Loss Prevention in the Process Industries, 41*, 399–410.

Berger, J. B., & Milem, J. F. (1999). The role of student involvement and perceptions of integration in a causal model of student persistence. *Research in Higher Education, 40*(6), 641–664.

Bouzembrak, Y., & Marvin, H. J. (2016). Prediction of food fraud type using data from Rapid Alert System for Food and Feed (RASFF) and Bayesian Network Modelling. *Food Control, 61*, 180–187.

Burtner, J. (2005). The use of discriminant analysis to investigate the influence of non-cognitive factors on engineering school persistence. *Journal of Engineering Education, 94*(3), 335.

Chen, R., & DesJardins, S. L. (2010). Investigating the impact of financial aid on student dropout risks: Racial and ethnic differences. *The Journal of Higher Education, 81*(2), 179–208.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research, 3*(Nov), 507–554.

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory, 14*(3), 462–467.

Colladon, A. F., & Remondi, E. (2017). Using social network analysis to prevent money laundering. *Expert Systems with Applications, 67*, 49–58.

Dag, A., Topuz, K., Oztekin, A., Bulur, S., & Megahed, F. M. (2016). A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. *Decision Support Systems, 86*, 1–12.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice, 13*(1), 17–35.

Donate, J. P., Cortez, P., Sanchez, G. G., & De Miguel, A. S. (2013). Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble. *Neurocomputing, 109*, 27–32.

Fallahi, A., & Jafari, S. (2011). An expert system for detection of breast cancer using data preprocessing and bayesian network. *International Journal of Advanced Science and Technology, 34*, 65–70.

Felten, P., Gardner, J. N., Schroeder, C. C., Lambert, L. M., Barefoot, B. O., & Hrabowski, F. A. (2016). *The undergraduate experience: Focusing institutions on what matters most*. John Wiley & Sons.

Forsman, J., Van den Bogaard, M., Linder, C., & Fraser, D. (2015). Considering student retention as a complex system: A possible way forward for enhancing student retention. *European Journal of Engineering Education, 40*(3), 235–255.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*: 1. New York, NY, USA: Springer series in statistics.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1–22.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*(2–3), 131–163.
Golde, C. M. (2005). The role of the department and discipline in doctoral student attrition: Lessons from four departments. *The Journal of Higher Education, 76*(6), 669–700.
Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student dropout predictive model using data mining techniques. *IEEE Latin America Transactions, 13*(9), 3127–3134.
Howard, J. S., & Flora, B. H. (2015). A comparison of student retention and first year programs among liberal arts colleges in the mountain south. *Journal of Learning in Higher Education, 11*(1), 67–84.
Hu, Y., Zhang, X., Ngai, E., Cai, R., & Liu, M. (2013). Software project risk analysis using Bayesian networks with causality constraints. *Decision Support Systems, 56*, 439–449.
Ishitani, T. T. (2016). Time-varying effects of academic and social integration on student persistence for first and second years in college: national data approach. *Journal of College Student Retention: Research, Theory & Practice, 18*(3), 263–286.
Johnson, N. (2012). *The institutional costs of student attrition*. Washington, D.C.: Delta Cost Project at the American Institute for Research Retrieved fromhttp://www.deltacostproject.org/resources/pdf/Delta-Cost-Attrition-ResearchPaper.pdf .
Käser, T., Klingler, S., Schwing, A. G., & Gross, M. (2017). Dynamic Bayesian networks for student modeling. *IEEE Transactions on Learning Technologies, 10*(4), 450–462.
Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT Press.
Korb, K. B., & Nicholson, A. E. (2010). *Bayesian artificial intelligence*. CRC press.
Lauría, E. J., Moody, E. W., Jayaprakash, S. M., Jonnalagadda, N., & Baron, J. D. (2013). Open academic analytics initiative: Initial research findings. In *Paper presented at the proceedings of the third international conference on learning analytics and knowledge*.
Lee, Y., & Choi, J. (2013). A structural equation model of predictors of online learning retention. *The Internet and Higher Education, 16*, 36–42.
Lin, S.-H. (2012). Data mining for student retention management. *Journal of Computing Sciences in Colleges, 27*(4), 92–99.
Lucas, P. J., van der Gaag, L. C., & Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine, 30*(3), 201–214.
Meyfroidt, G., Güiza, F., Ramon, J., & Bruynooghe, M. (2009). Machine learning techniques to examine large patient databases. *Best Practice & Research Clinical Anaesthesiology, 23*(1), 127–143.
Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. *Expert Systems with Applications, 38*(12), 14984–14996.
Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569.
Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.
Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Publishing Company, Incorporated.
Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems, 116*(8), 1678–1699.
Oztekin, A., Delen, D., & Kong, Z. J. (2009). Predicting the graft survival for heart-lung transplantation patients: An integrated data mining methodology. *International Journal of Medical Informatics, 78*(12), e84–e96. doi:10.1016/j.ijmedinf.2009.04.007.
Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96–146.
Pike, G. R., Hansen, M. J., & Childress, J. E. (2014). The influence of students' pre–college characteristics, high school experiences, college expectations, and initial enrollment characteristics on degree attainment. *Journal of College Student Retention: Research, Theory & Practice, 16*(1), 1–23.
Pittman, K. (2008). *Comparison of data mining techniques used to predict student retention* Ph.D. thesis. Nova Southeastern University.
Sevim, C., Oztekin, A., Bali, O., Gumus, S., & Guresen, E. (2014). Developing an early warning system to predict currency crises. *European Journal of Operational Research, 237*(3), 1095–1104.
Shih, D., Kim, S., Chen, V. P., Rosenberger, J., & Pilla, V. (2014). Efficient computer experiment-based optimization through variable selection. *Annals of Operations Research, 216*(1), 287–305. doi:10.1007/s10479-012-1129-y.
Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*(04), 687–719.
Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Procedia Computer Science, 1*(2), 2811–2819.
Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41*(2), 321–330.
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological, 58*(1), 267–288.
Tinto, V. (1997). Classrooms as communities: Exploring the educational character of student persistence. *The Journal of Higher Education, 68*(6), 599–623.
Tinto, V. (2012a). *Completing college: Rethinking institutional action*. University of Chicago Press.
Tinto, V. (2012b). Enhancing student success: Taking the classroom success seriously. *The International Journal of the First Year in Higher Education, 3*(1), 1–8.
Topuz, K., Uner, H., Oztekin, A., & Yildirim, M. B. (2017). Predicting pediatric clinic no-shows: A decision analytic framework using elastic net and Bayesian Belief Network. *Annals of Operations Research*, 1–21. doi:10.1007/s10479-017-2489-0.
Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., & Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems, 106*, 97–109.
Wang, G., Xu, T., Tang, T., Yuan, T., & Wang, H. (2017). A Bayesian network model for prediction of weather-related failures in railway turnout systems. *Expert Systems with Applications, 69*, 247–256.
Wang, Y., & Beck, J. (2013, July). Class vs. student in a Bayesian network student model. In *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 151–160). Berlin, Heidelberg: Springer.
Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and E-learning, 17*(1), 118–133.
Zhang, G., Anderson, T. J., Ohland, M. W., & Thorndyke, B. R. (2004). Identifying Factors Influencing Engineering Student Graduation: A Longitudinal and Cross-Institutional Study. *Journal of Engineering Education, 93*(4), 313–320.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.