Emerald Insight

Industrial Management & Data Systems

A Hybrid Data Analytic Approach to Predict College Graduation Status and its Determinative Factors

Asil Oztekin

A Hybrid Data Analytic Approach to Predict College Graduation Status and its Determinative Factors

Abstract

Purpose – The prediction of graduation rates of college students has become increasingly important to colleges and universities across the nation in the USA and across the world. Graduation rates, also referred to as completion rates, directly impact university rankings and represent a measurement of institutional performance and student success. In recent years, there has been a concerted effort by federal and state governments to increase the transparency and accountability of institutions, making "graduation rates" an important and challenging university goal. In line with this, the main target of this study is to propose a hybrid data analytic approach which can be flexibly implemented not only in the USA but also at various colleges across the world which would help predict the graduation status of undergraduate students due to its generic nature. It is also aimed at providing a means of determining and ranking the critical factors of graduation status.

Design/methodology/approach – This study focuses on developing a novel hybrid data analytic approach to predict the degree completion of undergraduate students at a four-year public university in the U.S. Via the deployment of the proposed methodology, the data is analyzed using three popular data mining classifications methods (i.e. decision trees, artificial neural networks, and support vector machines) to develop predictive degree completion models. Finally, a sensitivity analysis is performed to identify the relative importance of each predictor factor driving the graduation.

Findings – The sensitivity analysis of the most critical factors in predicting graduation rates are determined to be *fall term GPA, housing status (on campus or commuter),* and *which high school the student attended*. The least influential factors of graduation status are *ethnicity*, whether or not a student had *work study*, and whether or not a student applied for *financial aid*. All three data analytic models yielded high accuracies ranging from 71.56 % to 77.61%, which validates the proposed model.

Originality/value – This study presents uniqueness in that it presents an unbiased means of determining the driving factors of college graduation status with a flexible and powerful hybrid methodology to be implemented at other similar decision making settings.

Keywords: graduation rate, decision analytics, data mining, variable importance ranking

1. Introduction

1.1 Motivation

College graduation rates have become a primary focus in measuring institutional performance and accountability in higher education. In 1990, the Student Right to Know Act began requiring institutions for the first time to report graduation rates. Since then, graduation rates have become readily available to consumers and policymakers elevating the awareness and importance of graduation rates. The motivation from government stems not only from a consumer disclosure standpoint, but also from the fact that there is a significant benefit to society as a whole when a student graduates with a postsecondary degree. For example, postsecondary graduates are more likely to earn higher wages, carry health coverage, be civically engaged, and less likely to become unemployed than a person without a degree. From a more practical point of view, the legislators and policymakers who oversee higher education and allocate funds, the parents who pay for their children's education in order to prepare them for a better future, and the students who make college choices look for evidence of institutional quality and reputation to guide their decision-making processes, which is closely entailed with the graduation rates the higher education institutions (Gansemer-Topf and Schuh, 2006).

In 2009, President Obama set a goal for the United States to have the highest proportion of college graduates in the world by 2020. Based on a study conducted by the Organization for Economic Cooperation and Development (OECD), *24/7 Wall St*, the United States currently falls behind Canada, Israel, and Japan (Sauter & Hess, 2014). The President's commitment has put federal and state policymakers to the task to make colleges more accessible, affordable, and attainable. As a result, increased awareness and accountability is being placed on institutions to demonstrate performance at a higher level.

Beyond news articles and research studies, consumers today have access to sophisticated web tools to assist in their decision making process on which institution to attend. The College Navigator, developed by the National Center for Education Statistics (NCES), is a free consumer tool designed to aid students and families in obtaining institutional information such as programs, retention rates, graduation rates, and net costs. The Department of Education launched a similar tool for consumers in early 2013 called the College Scorecard. Although the purpose of this product is to provide transparency to the net price of institutions, it is yet another place where graduation rates are prominently displayed and compared nationally for each institution.

Also, released in 2013 by the Department of Education is the Financial Aid Shopping Sheet intended to standardize financial aid award notifications across institutions for students and families to easily compare net prices. In addition to the financial aid information provided, the institutions' median loan debt, cohort default rate, and completion rate are presented and compared to national averages on the sheet. A recent study showed that when a parent helps their child choose between colleges, the parent is more likely to select the college with the highest completion rate (Schneider and Kelly, 2014). With the heightened focus on transparency and accountability in higher education today, university administrators are developing internal strategies to improve graduation rates.

Increasing college graduation rates is not only important to institutions, but it is also significantly important to individuals and to the nation as a whole. Students who earn a postsecondary education obtain a variety of personal, financial, and long-term benefits (Baum *et al.*, 2010). An individual with a bachelor's degree will earn nearly twice as much as a worker with only a high school diploma (U.S. Department of Education, 2006). The unemployment rate for those with a bachelor's degree is about half of the unemployment rate for high school graduates. In addition, individuals with bachelor's degrees are more likely to have healthcare and pension benefits, be active citizens, and lead healthier lifestyles overall (Baum *et al.*, 2010). Clearly gainfully employed individuals contribute to the society overall. Higher earnings result in higher federal and state tax revenue. Lower unemployment rates reduce the monetary resources required to support public assistant programs. Increasing graduation rates will only increase these benefits to individuals and to society.

As the increase of college graduation rate has become a national agenda, the prediction of graduation rates has never been more important for institutions. In this study, the primary goal is to identify the most important predictors of graduation to assist universities to maintain or increase graduation rates. In addition, this study determines which indicators reduce the likelihood of graduation in order to identify pockets of at-risk students who may require additional resources to succeed.

1.2 Literature Review

Four-year institutions are most commonly measured by the six-year graduation rate. This graduation rate measures the six-year completion for only those students who enroll in and graduate from the same institution. Transfer students are not included in this calculation. Previous research has not yet determined an appropriate method to track students switching between institutions to calculate persistence and graduation rates of transfer students uniformly across institutions. According to the National Center for Education Statistic (NCES), approximately 58% of first-time students who began enrollment in a bachelor's degree at a four-year institution in fall 2004 earned this degree within six years from the institution at which they began their studies. The breakdown by the type of institution is 65% at private non-profit institutions, 56% at public institutions, and 28% at private for-profit institutions (Aud *et al.*, 2012).

College completion rates differ considerably by family income, parental education level, and type of institution attended (Baum *et al.*, 2010). Students whose parents have a bachelor's degree are more likely to complete their degree within six years. Likewise, students from higher income families are more likely to complete their degree within six years (Baum *et al.*, 2010). Recent research also suggests students who declare a major earlier in their undergraduate career are more likely to persist and complete their degree compared to those students who declare a major later on (Jenkins & Cho, 2012). Similarly, another study investigated as to when the student applied to the college. The earlier the student applied, the more likely they were to complete and receive their bachelor's degree (Hughes, 2012).

Student retention rates play an integral role in the completion agenda. An institution must first retain students in order to get them to degree completion. Delen (2010) performed a study on first-year retention using five years' worth of data from a large public university with an average enrollment of 23,000 students. About 80% of the population represented in state residents and approximately 19% were classified as minority. This study utilized the popular CRISP-DM data mining method and examined three different classification methods i.e. artificial neural networks, decision trees, and support vector machines. In this study, financial support and academic performance (past and present) were the most important predictors for student attrition. A related study performed by Higher Education Research Institute (HERI) also reported the accuracy of prediction increase with the addition of more variables (DeAngelo *et al.*, 2011).

Another similar research study was comprised of data from the Consortium for Student Retention Data Exchange (CSRDE), which is a voluntary consortium of four-year public and private post-secondary institutions (Hosch, 2011). The data used represented the first time, full-time freshmen Fall 2011 cohorts of these institutions, and the study utilized several models to identify the most important factors connected to student retention and ultimately graduation. The methodology used to complete this was linear regression. Through linear regression they focused on the factors that had a high correlation to whether or not the students completed their degree. The factors determined to be the most significant to the prediction of the six-year graduation rate were admission scores, the percentage of students living on campus, and first-semester college grade point average. This study reaffirmed similar results from earlier research conducted by Astin & Oseguera (2005) and Horn (2006) which demonstrated a strong correlation between admission scores and high school performance with retention and graduation rates.

Aside from the previous studies, there are also several other studies conducted in relation to completion rates. A regression analysis performed by Engle and O'Brien (2007) revealed that important predictive factors and models did not always produce the same level of accuracy across institutions. Their research demonstrated that some institutions performed better than expected while others performed worse than expected. Even though the key predictive factors of student characteristics were identified from the onset, it did not always help the college or university as planned. The gap in performance was attributed to the varying degree of academic support provided by the different institutions in facilitating students to persistence.

As though conventional statistical techniques such as logistic regression have been used in predicting student graduation status (Zhang *et al.,* 2004), data mining techniques have only recently gained popularity in the field of education. To exemplify, Kardan *et al.* (2013) used a powerful data mining method i.e. artificial neural networks in order to predict student behavior when selecting classes in an online higher education institute. Their study not only revealed patterns in class selection, but also in online student behavior and decision making processes. In a similar vein, Luft *et al.* (2013) also used artificial neural networks in order to estimate mathematical ability in an artificial game environment. Ultimately, their model successfully found patterns within the data, and was able to correctly predict the ability and achievement of individual mathematics students. Mendez and Gonzalez (2013) used fuzzy logic and artificial neural networks to measure and predict the in-class participation and assignment performance of engineering students. Their results yielded a system model based on Reactive Blended Learning, which incorporates both the teacher and student as crucial elements of participation and performance metrics. Feldman *et al.* (2014) deployed similar machine learning techniques such as neural networks, decision trees, and genetic algorithms in order to better understand the different ways in which students perceive and adopt new knowledge. Although the particular interest of this current study is data analytic modeling in education, there are several studies to exemplify the applications of data mining in various other fields, which has recently inspired the researchers to implement data mining in education domain as well (Lee&Siau, 2001).

2. Methodology

The popular CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is utilized in this study. CRISP-DM provides a comprehensive and systematic way of conducting data mining research, thereby improving the likelihood of obtaining accurate and dependable results. The model delivers a six-step process as summarized next (Shearer, 2000):

The first three steps (business understanding, data understanding, and data preparation) account for the vast majority of the total project time. Although each stage generates inputs for the next stage, the sequence is not strict and hence moving back and forth is actually required to achieve favorable results. A pictorial representation of the CRISP-DM is depicted in Figure 1.

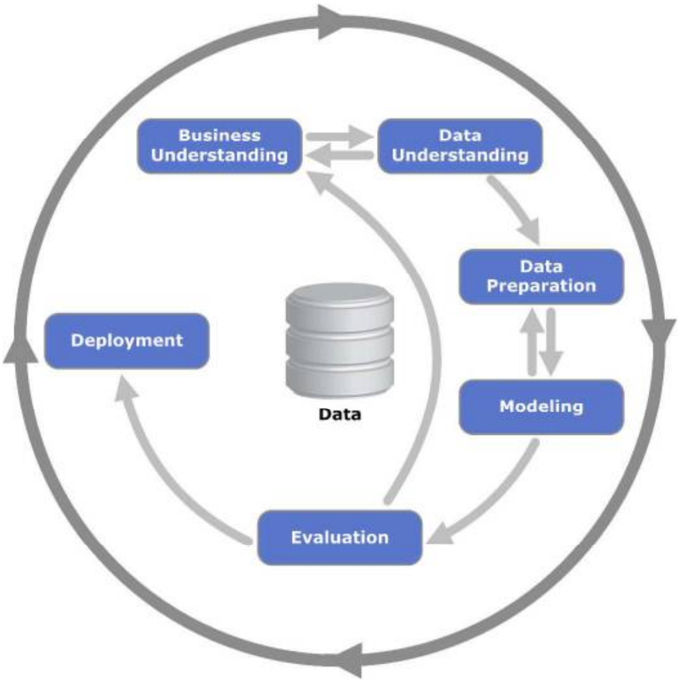

Figure 1. CRISP-DM process model

Once the initial source of data was identified for this research, the data was carefully examined for relevance and quality level. A significant amount of time was spent preprocessing the data into final dataset used for modeling. The data was scrubbed (i.e. cleansed) to remove missing values, exclude irrelevant records, and transform variables into appropriate descriptors for predictive modeling. Next the dataset was processed via data analytical models to predictive whether or not a student would graduate on time i.e. within six years. For each classification model, the data was separated into the 10 mutually exclusive data folds using the *k*-fold cross validation method. All folds except one were used to train each of the predictive classification models. The final fold was reserved to actually test the developed model and assess the baseline accuracy rate of its prediction. In this study, three different classifications models were used: decision trees, support vector machines, and artificial neural networks. If the predictive model does not yield an accuracy more than the random chance of 50% for this two-class (graduated on time or not) prediction, then it is not considered as a "satisfactory model" and hence is not deployed for subsequent analyses. If it does, then it is utilized for variable ranking via sensitivity analysis with respect to its accuracy score.

Once baseline accuracy rates were derived from the classification models, a sensitive analysis was conducted for each of the 30 input variables for all 3 models (i.e. the executable ran 90 times) in order to determine the importance order of variables to each of the models. Through the process of information fusion-based sensitivity analysis, the results from each variable were weighted against the three models and finally ranked in order of importance to the classification prediction. A flow chart of the methodology described in this section is shown in Figure 2.

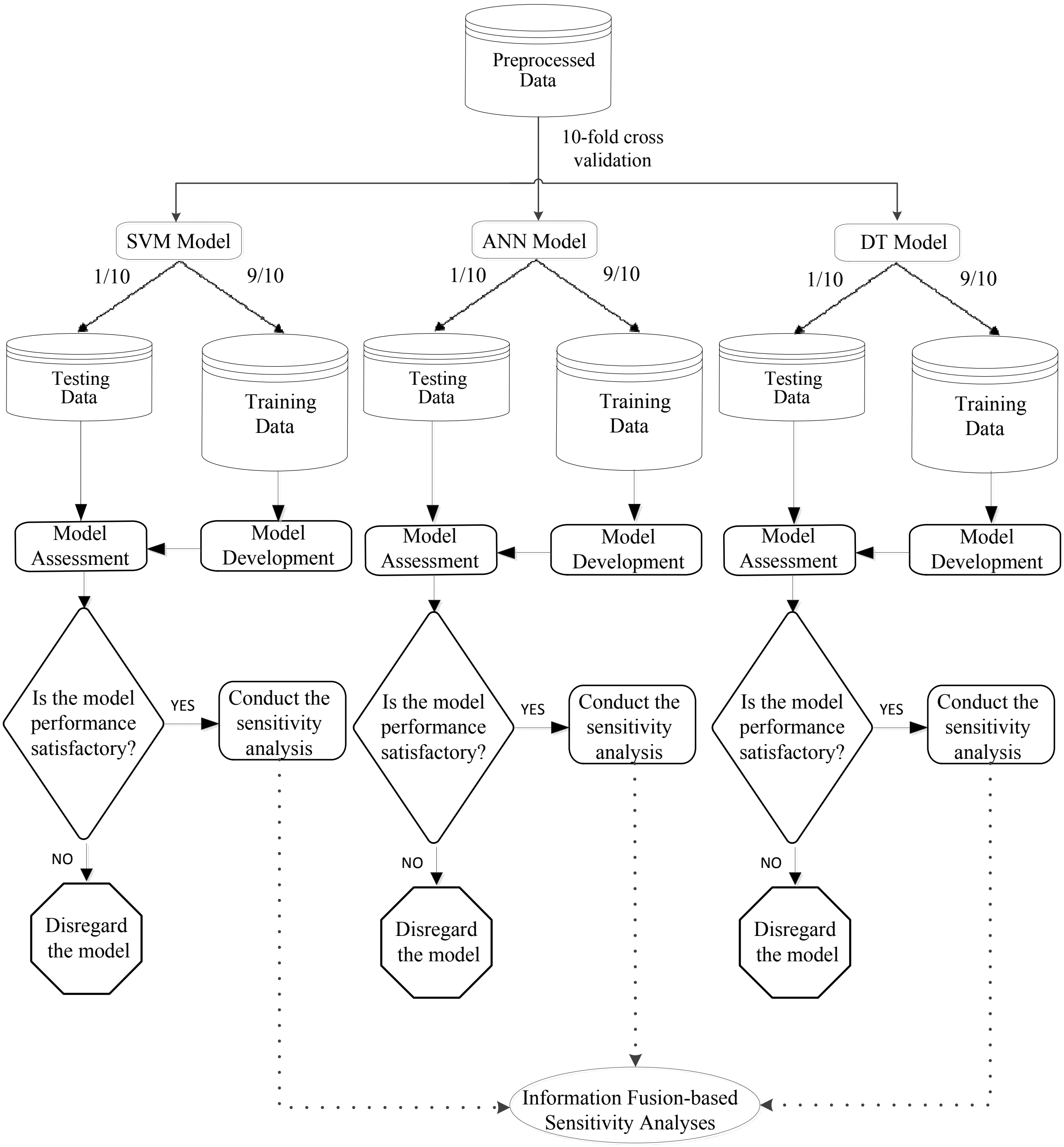

Figure 2. The proposed hybrid data analytic method

2.1 Data Analytic Models

Three data mining models (decision trees, artificial neural networks, and support vector machines) are employed in the proposed hybrid approach due to their popularity in literature which stems from the fact that they have conventionally outperformed many other methods in terms of accuracy (Delen *et al.*, 2010; Delen *et al.*, 2012; Oztekin *et al.*, 2009; Oztekin *et al.*, 2011; Oztekin, 2011; Oztekin, 2012; Oztekin *et al.*, 2013; Oztekin & Khan, 2014; Sevim *et al.*, 2014; Turkyilmaz *et al.*, 2013). Moreover, our trial-and-error experiments with these models also yielded superior results compared to the other potential prediction models such as logistic regression. Therefore, they have been reported in this study. What follows next is a brief summary of these data analytic models.

2.1.1 Decision Trees

Decision trees (DT) are one of the most understandable and easy to interpret prediction methods. This is one of the main reasons why they have gained popularity in several applications of analytics-related studies. The decision tree construction procedure starts with splitting the entire dataset into several subsets, which contain more or fewer homogeneous states of dependent variable (Turban *et al.*, 2010). At each split in the tree, the impacts of all predictor variables on the dependent variable are evaluated. This procedure takes place successively, until a decision tree gets to a stable state. Decision trees are powerful classification tools used in data mining, also offer a simplistic visual that can be explained easily to administrators and executives (Turban *et al.*, 2010).

Among the decision tree models, “classification tree” and “regression tree” present solutions to two different prediction modeling problems. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 (Quinlan, 1986; Quinlan, 1993) and C&RT (a.k.a CART) (Classification and Regression Trees) (Olshen *et al.*, 1984). Quinlan developed C4.5 and C5.0, in which the latter provides various superiorities over the former in terms of its efficiency and effectiveness (Quinlan, 1993). In other words, C5.0 is much more rapid than C4.5 which refers to its efficiency in terms of the computational time. It is also more “memory efficient” than C4.5. With regard to its effectiveness, it develops a fairly more compact tree while sustaining the same level of accuracy or sometimes even yields a superior one. Moreover, it provides more flexibility in that it allows for a different weighting scheme for variables/attributes while considering differing misclassification types. The last but not the least, it enables extracting the noise inherent in the data, which in turn prevents over-fitting, brings robustness and unbiasedness.

On the other hand, C&RTs were first introduced by Breiman *et al.* (1984), which is a binary DT algorithm. It can handle both numeric and non-numeric (categorical) variables. In other words, C&RT is capable of developing decision trees both for *classification* and *regression* type prediction problems, in which the output variable is a binary and continuous variable, respectively. This is what it significantly differentiates from C4.5 and C5.0 algorithms. C&RT first divides the data into two groups to render the records in each group more homogeneous so that all group records are in the same class value. Those two groups are then partitioned again until either some homogeneity criterion or other stopping criterion (e.g. number of epochs or computational time) is satisfied. In developing the C&RT decision tree, it is permissible to utilize the same independent attribute. The rationale behind splitting is to identify the best attribute related to the right threshold in order to ensure the optimal homogeneity of the subsets i.e. branches of the DT. The C&RT algorithm has been employed in this study due to its favorable performance compared to the other decision tree algorithms obtained in the preliminary analysis.

2.1.2 Artificial Neural Networks

Artificial neural networks (ANNs) are interconnected assembly of simple parallel processing elements, units or nodes, whose functionality is loosely based on an animal neuron. The processing ability of the network is stored in the inter-unit connection strengths or weights, obtained by a process of adaptation to or learning from a set of training patterns (Hassoun, 1995). Artificial neurons are the processing elements of ANNs. The very first step in ANNs is the computation of the weighted sums of all input elements entering each processing element (neuron). The net input of neuron *j* is

$$Y_j = \sum_i w_{ij} x_i + \theta_j \tag{1}$$

where x_i's are the outputs of the neurons in the previous layer, w_{ij} is the synaptic weight of neuron *i* to neuron *j*, and θ_j is the bias which is the constant value of the sigmoid function (Turban *et al.*, 2010). Then the weighted sum passes through a transformation (transfer) or activation function and this value becomes the output of the neuron. The Sigmoid logical activation function (or sigmoid transfer function) as given in Eq. (2)

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

is an S-shaped transfer function in the range of 0 to 1 and is a popular one used as an activation function because of its capability of capturing the nonlinearity (Turban *et al.*, 2010).

The most critical step in ANN is the training. Back propagation (BP), arguably is the most popular learning algorithm, is a gradient descent algorithm that propagates the errors through the network and allows adaptation of the hidden neurons (Principe *et al.*, 2001). It minimizes the total error via adjusting the weights along its gradient (Principe *et al.*, 2001). Root Mean Square Error (RMSE) is the conventionally used total error value, which can be calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - O_t)^2} \tag{3}$$

where O_t is the output vector and *t* is the index for output units . During BP learning, weights are modified according to their contribution to the error function, which is given by Eq. (4)

$$\Delta w_{ij} = -\eta \frac{\partial(RMSE)}{\partial w_{ij}} \tag{4}$$

where η is the learning rate which determines the magnitude of changes to be made in the learning parameter (Haykin, 2008).

Multilayer perceptron (MLP) models are the most popular neural network models. This type of neural network consists of a feed-forward network and supervised learning that requires a desired output in order to learn. In MLP models, there is always an input layer with a number of neurons equal to the number of variables of the problem, and an output layer, where the perceptron response is made available, with a number of neurons equal to the desired number of quantities computed from the inputs (Kizilaslan & Karlik, 2009).

2.1.3 Support Vector Machines

SVMs are one of the most popular machine learning methods based on Vladimir Vapnik's statistical learning theory (Cortes & Vapnik, 1995). SVM is a supervised learning algorithm, which may perform regression (hence sometimes called support vector regression) or classification using priori defined categories (Vapnik, 1998). Due to its useful features and promising empirical performance, SVM algorithm is gaining more popularity (Cho *et al.*, 2005).

Its structural risk minimization (SRM) feature shows superiority to other traditional empirical risk minimization (ERM)-based methods. SRM minimizes the expected risk of an upper bound while ERM minimizes the error of the training data. Hence, SVMs provide a good generalization performance with a computational efficiency in terms of speed and complexity, and easily deals with multi-dimensional data (Cho *et al.*, 2005). In addition, SVM works well under many circumstances even when there is a small sample dataset (Cristianini & Shawe-Taylor, 2000).

SVM classifier takes the inputs from different classes, and then builds input vectors into a feature space. An SVM model splits the training examples into separated categories in a mapped space as certain points (Kecman, 2005). SVM is to map these points as vectors into a higher dimensional feature space. The vectors transform a linear or non-linear map into the feature space (Shiue, 2009). SVMs use an optimal hyperplane to separate classes in a data set. A hyperplane which places at the maximum distance from the nearest points of the data set is defined as the optimal (Cortes & Vapnik, 1995).The points which determine optimal hyperplane to separate different classes in a data set are called Support Vectors. These are critical elements to train the classifying algorithm (Kecman, 2005).

In a feature space, to find the hyperplane, Lagrange multipliers are introduced to solve a quadratic problem. SVM's classification process is briefly reviewed as follows.

x_i is a feature of the i^{th} example, y_i denotes an output for the i^{th} example as a binary value; w is a weight and b is a bias, following inequalities in Eq. (5) and Eq. (6) can be written.

$$wX_i + b \geq +1 \, for \; y_i = +1 \tag{5}$$

$$wX_i + b \leq -1 \, for \; y_i = -1 \tag{6}$$

When conditions 5 and 6 are considered for all pairs of (x_i, y_i) for i=1, 2… m, expression (7) would be followed.

$$\{(x_i, y_i) | x_i \in R^N, y_i \in \{-1,1\}\}_{i=1}^{m} \tag{7}$$

where m labels are the given samples in a data set.

The optimal hyperplane is achieved by maximizing a margin between support vectors where $||w||$ denotes the maximization of the margin by following a quadratic problem solution.

$$\min \frac{1}{2} \|w\|^T \|w\|$$
$$s.t.\ \ y_i(w^T x_i + b) \geq 1 \ and\ i = 1,2,\dots,m \qquad (8)$$

Karush-Kuhn-Tucker conditions apply to this problem as shown in Eq. (9).

$$\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$
$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C \ and\ i = 1,2,\dots,m \qquad (9)$$

where *C* is a penalty parameter, using Lagrange multipliers assist to solve the problem (Cristianini & Shawe-Taylor, 2000).

SVM algorithms also use kernels to reduce the complexity of problems as mapping them in a higher dimensional space (Vapnik, 1998). If the data is not separable, using a kernel trick it becomes easier to classify the inputs. The kernel function allows mapping the same data in a linearly separable way on a higher dimensional space. The trick is formed as in function *K*.

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \qquad (10)$$

Polynomial kernel, Gaussian radial basis function (as in Eq. (11)), and sigmoid function (as in Eq. (12)) are among the most commonly used kernel functions.

$$K(x_i, y_j) = \left((x_i, y_j) + 1\right)^p \qquad (11)$$
$$K(x_i, y_j) = e^{(-\|(x_i - y_j)\|^2 / 2\sigma^2)} \qquad (12)$$

As a kernel function is imposed, the problem transforms into a new quadratic model as shown in Eq. (13).

$$max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j \varphi(x_i)^T \varphi(x_j)$$

$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \ and \ i = 1,2, \dots, m \tag{13}$$

Finally, a decision, which is not about a magnitude but a sign, is made as shown in Eq. (14):

$$D(x) = sign \ (\textstyle\sum_{k=1}^{n} a_k y_k K(x, x_k) + b) \tag{14}$$

More detailed information about SVM algorithms and applications can be found in Vapnik (1998), Cristianini & Shawe-Taylor (2000), and Kecman (2005).

2.2 Performance Measures

2.2.1 *k*-Fold Cross Validation

The *k*-fold cross validation, also known as the rotation estimation, randomly splits the full dataset into *k* mutually exclusive data subsets of approximately equal size (Kohavi, 1995). All but one subset is used in the training of model. Each of the subsets used for training are executed *k* number of times. Ten is a popular number used for the number of folds (*k*) (Kohavi, 1995). The remaining one data subset is used in the final testing of model. Results are aggregated for a true estimation of the prediction accuracy. The *k*-fold cross validation minimizes the bias that is often associated with random sampling of the training dataset. The reasoning of the choice for *k*=10 is based on literature (Kohavi, 1995) which shows that 10-folds provide an ideal balance between the classification performance and the time required to run the models. Additional details on *k*-fold cross validation can be found in Han *et al.* (2011). A graphical representation of 10-fold cross validation is illustrated as in Figure 3.

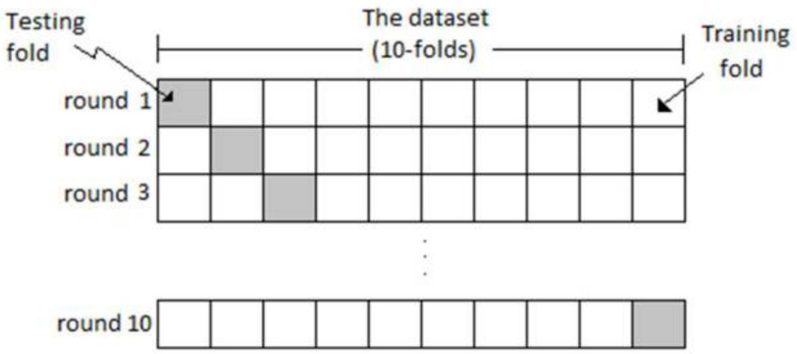

Figure 3. A pictorial representation of 10-fold cross-validation

2.2.2 Confusion matrix

A confusion matrix is a tabular representation of prediction (specifically two-class classification) model outcomes. The rows represent the actual classes and the columns represent the predicted classes. For example, a two-class prediction (i.e. classification) problem display results as follows:

True positive: number of samples classified as true while they were true
False positive: number of samples classified as true while they actually were false
False negative: number of samples classified as false while they actually were true
True negative: number of samples classified as false while they were actually false

The performance metrics for the classification models performed in this study are accuracy, sensitivity, and specificity are explained as in Eq. (15), Eq. (16), and Eq. (17), respectively:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

$$Specificity = \frac{TN}{TN + FP} \tag{17}$$

Accuracy measures the proportion of correctly classified examples to give an overall probability that the model can correctly classify. Sensitivity and specificity, shown by Eqs. (16) and (17), measure the model's ability to recognize a specific class. For example, sensitivity is the probability that a student will graduate on time (within 6 years) and specificity is the probability that a non-graduating student is predicted to be graduating on time, in fact, will not graduate.

2.2.3 Information Fusion-Based Sensitivity Analysis

In data mining, there is no single method that would work the best for each and every problem. In other words, the performance of a model is derived by the studied scenario and the dataset being utilized (Oztekin *et al.,* 2013*;* Sevim *et al.*, 2014). In a similar vein, it is impossible to state the best strategy to deploy various data mining methods. Therefore, in order to receive more accurate and effective results out of a set of data mining models, researchers tend to use "composite forecasts" stemming from the integration of multiple models (Oztekin *et al.,* 2013; Sevim *et al.*, 2014). The *information fusion* outlines a process of combining the information extracted from a set of data mining models. There is a consensus that such a fusion produces more useful information in knowledge discovery in databases (KDD) practices (Batchelor & Dua, 1995; Chase, 2000).

The information fusion algorithm can be as formulated as in Eq. (18) where the output (dependent) variable is shown by variable y and the input (independent) variables by $x_1, x_2, \ldots, x_n$ (Delen *et al.,* 2007).

$$\hat{y} = f(x_1, x_2, \ldots, x_n) \tag{18}$$

If m number of prediction models are employed, the fusion model can be written as in Eq. (19)

$$\hat{y}_{fused} = \psi(\hat{y}_{individual\ i}) = \psi(f_1(x), f_2(x), \ldots, f_m(x)) \tag{19}$$

where ψ the is the operator to fuse/integrate the predictions of models $f_1(x), f_2(x), \ldots, f_m(x)$.

If fusing operator ψ is a linear function, as the case in this study, then we can rewrite Eq. (19) as in Eq. (20).

$$\hat{y}_{fused} = \sum_{i=1}^{m} \omega_i f_i(x) = \omega_1 f_1(x) + \omega_2 f_2(x) + \cdots + \omega_m f_m(x) \tag{20}$$

where $\omega_1, \omega_2, \ldots, \omega_m$ refer to the weighting coefficients of each individual model, namely, $f_1(x), f_2(x), \ldots, f_m(x)$. Also, it can be assumed that the weights are normalized so that $\sum_{i=1}^{m} \omega_i = 1$ holds true.

The weights (ω's) are assigned proportional to the performance measure of each data mining model. In other words, the higher the accuracy of a data mining model, the higher the weight of that particular data mining model results (Oztekin *et al.,* 2013; Sevim *et al.*, 2014). Moreover, putting independent variables in rank order in terms of their importance in prediction is also critical. In Artificial Neural Networks, *sensitivity analysis* is the technique to do so for a trained ANN model (Davis, 1989). Through the sensitivity analysis, the learning algorithm of the ANN model is disabled after the learning is accomplished so that the network weights are not affected. Hence, the sensitivity score of a given input/independent variable is the percentage ratio of the ANN model error without the specified independent variable to the error of the model with all independent variables (Mitchell, 1997). The more the model deterioration is without the particular variable, the higher the importance level of that variable would be. The same philosophy is valid in SVMs to determine variable rank order in terms of their importance as well, which is quantified as "sensitivity measure" defined by Eq. (21) (Saltelli, 2002).

$$S_i = \frac{V_i}{V(F_t)} = \frac{V(E(F_t|X_{-i}))}{V(F_t)} \tag{21}$$

where, $V(F_t)$ is the unconditional output variance. In the numerator, the expectation operator *E* calls for an integral over X_{-i}; that is, over all input variables but X_i, then the variance operator *V* implies a further integral over X_i (Saltelli, 2002). The variable importance is then computed as the normalized sensitivity (Saltelli *et al.,* 2004).

Considering Eqs. (20) and (21) simultaneously, the sensitivity measure of the variable n with the information fused by m prediction models can then be established as given in Eq. (22)

$$S_{n(fused)} = \sum_{i=1}^{m} \omega_i S_{in} = \omega_1 S_{1n} + \omega_2 S_{2n} + \cdots + \omega_m S_{mn} \qquad (22)$$

where, ω's refer to the normalized performance measure (i.e. accuracy value) of each prediction model with m models in total; and S_{in} is the sensitivity measure of the n^{th} variable in the i^{th} model.

3. Data Source and Attribute Descriptions

The data used in this study is from a large four-year public university located in the U.S. It was acquired through official data acquisition and was completely anonymized when received from the corresponding institution in compliance with the confidentiality and privacy. Nearly 90% of students are in-state residents and enrolled in full-time study. The majority of students are commuters, and males represent over 50% of the student body. The institution's six year graduation rate is about 51%.

The dataset is comprised of students who entered the Fall 2007 semester as the first time freshman and was extracted from the institution's student relational database. The data variables include demographic characteristics, financial indicators, and academic qualities of each student.

The data was analyzed and cleansed to remove erroneous records and incomplete data elements. For example, students who did not actually begin attendance were removed from the dataset as were those students without SAT scores. In addition, 4 data variables were created from the original data and included in final dataset. These variables included *fall enrollment status, fall completion rate, spring completion rate,* and *six-year degree completion.*

The final dataset used in this study contained 1204 records and 31 variables, 30 used as predictors (input variables) and 1 used as the dependent (output) variable. A summary of data variable descriptions are provided in Table 1.

Table 1. Descriptions of variables

#	Variables	Data Type	# of Unique Values
1	College	Multi-nominal	5
2	Major	Multi-nominal	38
3	Age	Numeric	10
4	Gender	Binary nominal	2
5	Ethnicity	Multi-nominal	9
6	Citizenship	Multi-nominal	5
7	Residency status	Multi-nominal	5
8	High school	Multi-nominal	275
9	High school GPA	Numeric	220
10	SAT score	Numeric	86
11	SAT Math score	Numeric	52
12	SAT Verbal score	Numeric	52
13	Housing status	Binary nominal	2
14	Financial aid applicant	Binary nominal	2
15	Financial need type	Multi-nominal	6
16	Athletic scholarship	Binary nominal	2
17	Merit scholarship	Binary nominal	2
18	Grant - need based	Binary nominal	2
19	Loan - student	Binary nominal	2
20	Loan - parent	Binary nominal	2
21	Student employment	Binary nominal	2
22	Fall term GPA	Numeric	582
23	Fall enrollment status	Binary nominal	2
24	Fall completion rate	Numeric	40
25	Fall attempted credits	Numeric	20
26	Fall earned credits	Numeric	21
27	Spring term GPA	Numeric	514
28	Spring completion rate	Numeric	31
29	Spring attempted credits	Numeric	21
30	Spring earned credits	Numeric	21
31	Degree Completion- Six Year	Binary nominal	2

4. Case Study Results and Managerial Implications

In this study, with all 30 variables included, the accuracy, sensitivity, specificity, and precision values were calculated for each classification model and can be seen in Table 2.

Table 2. 10-fold cross-validated prediction results (in %)

Model Type	Accuracy	Sensitivity	Specificity
Decision Tree	73.75	64.23	82.09
Neural Network	71.59	66.19	76.32
Support Vector Machine	77.61	71.48	83.21
Logistic Regression	50.18	45.60	55.43

Using 10-fold cross-validation the support vector machine had the best results, followed by decision tree, and then neural networks. The support vector machine model had an accuracy of 77.61%, sensitivity of 71.48%, and specificity of 83.21%. The second best results were derived from the decision tree with an accuracy rate of 73.75%, sensitivity of 64.23%, and specificity of 82.09%. Finally, the classification model that provided lowest results was neural network. The accuracy for neural network model was 71.59%, the sensitivity was 66.19%, and the specificity was 76.32%. On the other hand, logistic regression which is a linear modeling technique and is dependent on strong statistical assumptions did not yield satisfactory results and therefore was not considered in the subsequent analyses of information fusion-based sensitivity scoring as depicted in Figure 2 earlier.

The input variables that were more sensitive to the output were able to be determined via the information fusion-based sensitivity analysis calculation as explained in Section 2.2.3. This methodology by deploying 10-fold cross-validated three different data analytic models revealed 3x10=30 different scores for variable importance ranking with respect to their contribution to the prediction of a student's graduation status in six years. This is in a way to *let the data speak* objectively in an unbiased way to determine the effective factors of student graduation status. As shown in Figure 4, fall term GPA was the predictor that was the most important when it came to predicting whether or not a student will complete the university within 6 years. The predictor that was the second most important was housing status, whether or not a student lived on campus or commuted. The tertiary important predictor was which high school the student attended. The predictors that were the least important were ethnicity, whether or not the student was employed, and whether or not the student applied for financial aid. With the use of data mining tools and methods, university administrators would be better equipped to predict those students who will graduate within six years, and possibly more importantly, those who will not. As explained previously, increasing the college graduation rate has become a national agenda. Institutions are expected to perform at higher level than ever before, and all in the public eye. Graduation rates are increasingly becoming a critical measurement of an institution's performance and success to which students are using in their decision making process when selecting a college to attend.

The data mining results in this study provide university administrators with a guide to predict graduation rates. The study revealed that performance in the first semester is the most significant factor in predicting whether or not a student will graduate within six years. Furthermore, it is much more indicative of success than the admissions-based criterion of SAT score and high school GPA. This suggests that it is far more prudent for administrators to focus efforts on first year services, such as first year seminars, to support student retention and success.

The data results also indicated that students living on campus were more likely to complete within six years. This hypothetically suggests that on-campus students are more likely to have a greater sense of community than those who commute. Many institutions struggle with involving commuting students at the same level as on-campus students. More interesting than the most important variables are the least important ones: loan, financial aid application, and student employment status. These input variables, all of which are related to monetary issues, were revealed to be of low significance to the prediction of students' graduation within 6 years. Likewise, it is very encouraging to reveal the fact that the very least important predictor is the ethnic background of the student. In this study, the predominant indicators were those attributes of student performance and characteristics once a student enrolled, such as fall GPA, housing status, and spring GPA. It warrants focus on support services for new students. However, it does not reveal pre-admission factors and characteristics that will assist institutions to predict graduation from applicants. Further research to evaluate only those attributes related to high school students (i.e. removing college-related attributes such as academic performance) may also be valuable to institutions in making *admission* decisions. In fact, in literature it was discovered that more selective universities do not necessarily have higher graduation rates (Hermanowicz, 2003). Instead, other factors which are not directly related to "selectivity" would hypothetically be influential. Those potential factors which are associated with the "cultural side" of the college, such as norm and values that guide communities, should receive equal attention because a higher rate of retention is often achieved when students find the environment in their university to be highly correlated with their interests (Hermanowicz, 2003). Housing status, which is found to be the second important factor in Figure 4 can easily be considered under this category of factors.

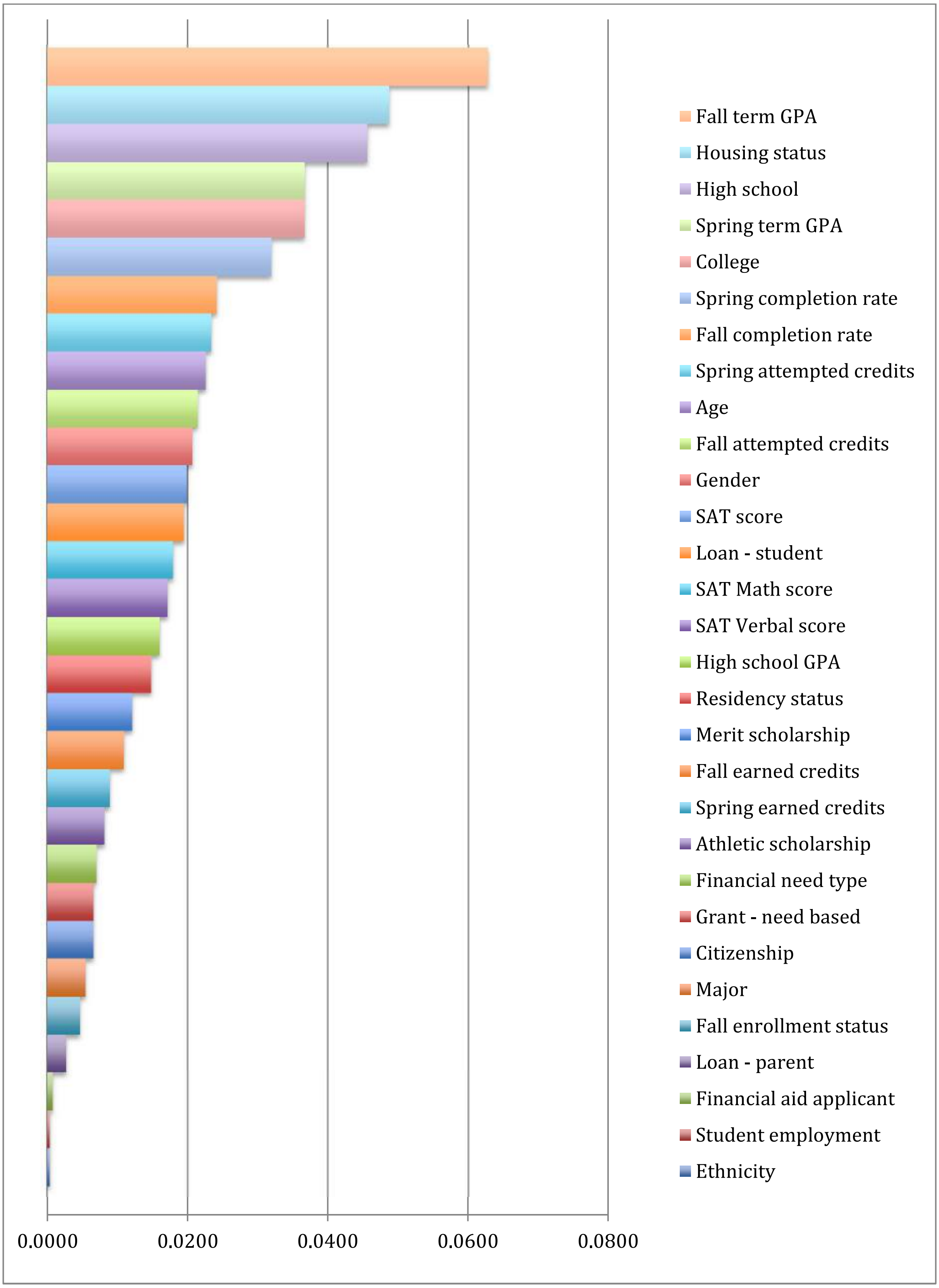

Figure 4. 10-fold cross-validated information fusion-based sensitivity analysis results

5. Conclusions

College students' on-time graduation has recently appeared to be one of the critical priorities for decision makers at higher education institutions. Improving graduation rates starts with a thorough understanding and analysis of the deriving factors behind. Such an understanding is the basis for accurately predicting at-risk students and appropriately intervening to graduate them on time. This study focused on developing a novel hybrid methodology for the prediction of degree completion of undergraduate students at a four-year public university using 3 different data analytic methods. To avoid the bias of random data splitting for training and testing, a 10-fold cross validation was utilized. In order to reveal the rank order of the predictor variables, an information fusion-based sensitivity analysis of each variable was performed. Once completing the classification models and utilizing the accuracy methods, the independent variable that mostly affected the prediction of degree completion was able to be determined. Through the use of data mining techniques, the results showed that fall term GPA, housing status, and high school were the independent variables that were the three highest determinative factors while monetary variables and the ethnic background of the student were revealed to be the least important ones. Referring to the findings as represented in Figure 4, university administrators at that particular institution should rather focus their efforts onto the admissions of the students since the high school where the students came from was a significant determinative factor of on-time graduation. Administrators of that college can develop better campaigning strategies to attract high performing high schools' students to this college, which historically have performed better than their peers. Also, another important issue to further emphasize should be developing more effective ways to promote on-campus stay of the students when they first get admitted to the college. This can be successfully achieved via a strong collaboration with the residential life office of that university. Alternatively, affordable and more flexible payment options for on-campus stay might be offered to the students which would help create an attractive option as opposed to commuting or other means of stay around the city the institution is. After the students are selected from better high schools and being admitted to the on-campus dorms and/or university apartments, then a close-watch strategy should be developed to track and trace students' grades. Timely notifications of missed classes both to the students themselves and to the parents/patrons might be an option to consider. Individual instructors whose students miss classes regularly should be in well-established communication with the problematic students. On the other hand, officially allocated "undergraduate advisors" who can closely follow up student success at each and every class s/he is enrolled in appears to be essential as well. Other more effective advising strategies could be developed to further improve students' Fall GPA since it turned out to be the most effective factor for graduation on time. Moreover, all the parties including faculty members, advisors, registrar's office, advising council and etc. at that institution should be well informed about the findings of such a study so as to make policy changes to improve their graduation rates. Training on the related parties would hypothetically be helpful. One interestingly striking result of this study is that in contrast to what is heuristically deemed to be very critical, i.e. monetary support is not that effective on the college graduation status, at least for the institution analyzed in this study. This might cause a paradigm shift in the way that the institution operates and efforts of the higher admins would need to be revised.

The data for this study was from the freshmen class that entered the analyzed institution in 2007. Since then, the institution has been undergoing many changes, these results might vary from year to year. For future research for this institution, the variables should be split up. One future study could only consist of variables before the students enter the college (i.e. high school GPA, SAT scores). The other study, for the same freshmen class, could only consist of variables after the students first year at the institution (i.e. cumulative college completion rate and GPA). With the separation of these variables, a study would be able to specifically determine the best classifiers before the student enters the institution and after the student completes a year at the institution. Through the use of data mining methods, these studies will be able to help the administrators, professors, and students increase the completion rate at the university by focusing on the important factors.

This study illustrates that the data mining prediction models can effectively predict students' degree completion within six years. The proposed methodology formulation is generic, though the case of an US-based institution has been used for illustration to highlight various aspects of the proposed framework. However, by using a different dataset for the course of interest, the procedure presented in this paper can be adopted for any given institution to determine the critical success factors in that area of study; and by analyzing the dataset for the given institution, guidelines can be developed that university administrators can adopt for improving their students' retention and enabling them to complete their degrees on time.

Acknowledgement

I am thankful to the three anonymous reviewers whose comments have helped significantly improve an earlier version of this paper. I am also grateful to the co-Editor-in-Chief of *Industrial Management & Data Systems* journal, Dr. Hing Kai Chan, for his very timely management of this manuscript.

References

Armstrong, J.S. (2002), *Combining Forecasts*, In J.S. Armstrong (Ed.), Principles of Forecasting, Kluwer Academic Publishers, Norwell, MA, pp. 418–439.

Astin, A.W. & Oseguera, L. (2002), *Degree Attainment Rates at American Colleges and Universities*, Los Angeles: Higher Education Research Institute, UCLA.

Aud, S., Hussar, W., Johnson, F., Kena, G., Roth, E., Manning, E., Wang, X., Zhang, J. (2012), "The Condition of Education", *U.S. Department of Education, National Center for Education Statistic (NCES) Washington, DC: U.S. Government Printing Office,* pp. 1-378.

Batchelor, R., and Dua, P. (1995), "Forecaster Diversity and The Benefits of Combining Forecasts," *Management Science,* Vol. 41, pp. 68–75.

Baum, S., Ma, J., Payea, K. (2010), "Education Pays 2010: The Benefits of Higher Education for Individuals and Society", *College Board Trends in Higher Education Series*, Vol. , pp. 1-52.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone C.J. (1984), *Classification and Regression Trees*, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chase Jr, W.C. (2000), "Composite Forecasting: Combining Forecasts for Improved Accuracy", *Journal of Business Forecasting Methods & Systems,* Vol. 19.

Cho, S., Asfoura, S., Onar, A., Kaundinya, N. (2005), "Tool breakage detection using support vector machine learning in a milling process", *International Journal of Machine Tools and Manufacture*, Vol. 45, pp. 241–249.

Cortes, C., Vapnik, V.N. (1995), "Support vector networks", *Machine Learning*, Vol. 20, pp. 273-297.

Cristianini, N., and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines* *and Other Kernel-based Learning Methods*, 1st Ed., Cambridge University Press, Cambridge, U.K.

Davis, G.W. (1989), "Sensitivity Analysis in Neural Net Solutions", *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 19, pp. 1078-1082.

DeAngelo, L., Franke, R., Hurtado, S., Pryor, J., Tran, S. (2011), *Completing college: Assessing graduation rates at four-year institutions*, Los Angeles: Higher Education Research Institute at the University of California, pp. 1-55.

Delen, D., Oztekin, A., Kong, Z.J. (2010), "A machine learning-based approach to prognostic analysis of thoracic transplantations", *Artificial Intelligence in Medicine*, Vol. 49 (1), pp. 33-42.

Delen, D., Oztekin, A., Tomak, L. (2012), "An analytic approach to better understanding and management of coronary surgeries", *Decision Support Systems* Vol. 52 (3), pp. 698-705.

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

Delen, D., Sharda, R., Kumar, P. (2007), "Movie Forecast Guru: A Web-Based DSS for Hollywood Managers", *Decision Support Systems,* Vol. 43, pp. 1151-1170.

Engle, J., O'Brien, C. (2007), *Demography Is Not Destiny: Increasing the Graduation Rates of Low-Income College Students at Large Public Universities*, The Pell Institute, pp. 1-68.

Feldman, J., Montreserin, A., Amandi, A. (2014), "Detecting student's perception style by using games", *Computers and Education*, Vol. 71, pp. 14-22.

Gansemer-Topf, A.M., Schuh, J.H. (2006), "Institutional selectivity and institutional expenditures: examining organizational factors that contribute to retention and graduation", *Research in Higher Education*, Vol. 47, pp. 613–642.

Han, J., Kamber, M., and Pei, J. (2011), *Data Mining Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publisher.

Hassoun, M.H. (1995*), Fundamentals of Artificial Neural Networks*, MIT Press, Massachusetts.

Haykin, S. (2008), *Neural Networks and Learning Machines*, 3rd edition, Prentice Hall, New Jersey.

Hermaniwicz, J.C. (2003). *College Attrition at American Research Universities: Comparative Case Studies*, Agathon Press, New York.

Horn, L. (2006), *Placing college graduation rates in context: How 4-year college graduation rates vary with selectivity and the size of the low-income enrollment*, National Center for Education Statistics, pp. 1-155.

Hosch, B. (2008), *Institutional and Student Characteristics that Predict Graduation and Retention Rates*, North East Association for Institutional Research Annual Meeting, pp. 1-13.

Hughes, K. (2013), *The College Completion Agenda: 2012 Progress Report*. The College Board Advocacy & Policy Center, pp. 1-35.

Jenkins, D., & Cho, S. W. (2012), *Get with the program: Accelerating community college students' entry into and completion of programs of study* (CCRC Working Paper No. 32). New York, NY: Columbia University, Teachers College, Community College Research Center.

Kardan, A., Sadeghi, H., Ghidar, S., Sani, M. (2013), "Prediction of student course selection in online higher education institutes using neural network", *Computers and Education*, Vol. 65, pp. 1-11.

Kecman, V. (2005), *Support Vector Machines: An Introduction in Support Vector Machines: Theory and Applications*, In: Eds: L. Wang, Berlin: Springer-Verlag, Ch.1, pp. 1-47.

Kizilaslan, R., and Karlik, B. (2009), "Comparison Neural Networks Forecasters for Monthly Natural Gas Consumption Prediction", *Neural Network World,* Vol. 19, pp. 191-199.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection", *In: Proceedings of the 14th International Conference on AI (IJCAI)*, San Mateo, CA: Morgan Kaufmann, pp. 1137–1145.

Lee, S. J., Siau, K. (2001), "A review of data mining techniques", *Industrial Management & Data Systems*, Vol. 101, pp. 41 – 46.

Luft, C., Gomes, J., Priori, D., Takase, E. (2013), "Using online cognitive tasks to predict mathematics low school achievement", *Computers and Education*, Vol. 67, pp. 219-228.

Mendez, J., Gonzalez, E. (2013), "A control system proposal for engineering education", *Computers and Education,* Vol. 68, pp. 266-274.

Mitchell, T. (1997), *Machine Learning*, McGraw-Hill, New York.

Olshen, L., Stone, C.J. (1984), *Classification and Regression Trees*, Wadsworth International Group.

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,* Vol. 78 (12), pp. e84-e96.

Oztekin, A., Kong, Z.J., Delen, D. (2011), "Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations", *Decision Support Systems*, Vol. 51 (1), pp. 155-166.

Oztekin, A. (2011), "A decision support system for usability evaluation of web-based information systems", *Expert Systems with Applications*, Vol. 38 (3), pp. 2110-2118.

Oztekin, A. (2012), “An Analytical Approach to Predict the Performance of Thoracic Transplantations”, *Journal of CENTRUM Cathedra: The Business and Economics Research Journal,* Vol. 5 (2), pp. 185-206.

Oztekin, A., Khan, R. (2014), “A Business-Analytic Approach to Identify Critical Factors in Quantitative Disciplines”, *Journal of Computer Information Systems*, Vol. 54 (4), pp. 60-70.

Oztekin, A., Delen, D., Turkyilmaz, A., and Zaim, S. (2013), “A machine learning-based usability evaluation method for eLearning systems”, *Decision Support Systems,* Vol. 56, pp. 63–73.

Principe, J.C., Euliano, N.R., and Lefebvre, W.C. (2001), *Neural and Adaptive Systems: Fundamentals through Simulations*, John Wiley & Sons, New York.

Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Quinlan, J.R. (1986), “Induction of decision trees”, *Machine Learning*, Vol. 1, pp. 81-106.

Saltelli, A. (2002), “Making Best Use of Model Evaluations to Compute Sensitivity Indices”, *Computer Physics Communications,* Vol. 145, pp. 280–297.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004), *Sensitivity Analysis in Practice – A Guide to Assessing Scientific Models*, John Wiley and Sons.

Sauter, M., Hess, A.E.M. (2014), “The most educated countries in the world”, *247 Wall St.com,* http://finance.yahoo.com/news/the-most-educated-countries-in-theworld.html?page=all

Schneider, M., Kelly, A. (2014), “What parents don't know about college graduation rates can hurt”, *American Enterprise Institute,* http://www.aei.org/article/education/higher-education/what-parents-dont-know-about-college-graduation-rates-can-hurt/

Sevim, C., Oztekin, A., Bali, O., Guresen, E. (2014), “Developing an Early Warning System to Predict Currency Crises”, *European Journal of Operational Research,* Vol. 237, pp. 1095–1104.

Shearer, C. (2000), "The CRISP-DM model: the new blueprint for data mining", *Journal of Data Warehousing*, Vol. 5, pp. 13–22.

Shiue, Y. (2009), "Data-mining-based dynamic dispatching rule selection mechanism for shop floor control systems using a support vector machine approach", *International Journal of Production Research*, Vol. 47, pp. 3669-3690.

Turban, E., Sharda, R., Delen, D. (2010), *Decision Support qnd Business Intelligence Systems,* (9th edition), New Jersey, USA: Pearson Prentice Hall.

Turkyilmaz, A., Oztekin, A., Zaim, S., Demirel, O.F. (2013) "Universal Structure Modeling Approach to Customer Satisfaction Index", *Industrial Management & Data Systems*, Vol. 113 (7), pp.932 - 949.

U.S. Department of Education. (2006), "A Test of Leadership: Charting the Fund of U.S. Higher Education". Washington D.C., pp. 1-55.

Vapnik, V.N. (1998), *Statistical Learning Theory: Adaptive and Learning Systems for Signal Processing*, *Communications, and Control*, John Wiley & Sons, New York.

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*, Vol. 93, pp. 313–320.