

Survival Analysis based Framework for Early Prediction of Student Dropouts

Sattar Ameri
Wayne State University
Detroit, MI - 48202
ameri@wayne.edu

Ratna B. Chinnam
Wayne State University
Detroit, MI - 48202
ratna.chinnam@wayne.edu

Mahtab J. Fard
Wayne State University
Detroit, MI - 48202
fard@wayne.edu

Chandan K. Reddy
Virginia Tech
Arlington, VA - 22203
reddy@cs.vt.edu

ABSTRACT

Retention of students at colleges and universities has been a concern among educators for many decades. The consequences of student attrition are significant for students, academic staffs and the universities. Thus, increasing student retention is a long term goal of any academic institution. The most vulnerable students are the freshman, who are at the highest risk of dropping out at the beginning of their study. Therefore, the early identification of “at-risk” students is a crucial task that needs to be effectively addressed. In this paper, we develop a survival analysis framework for early prediction of student dropout using Cox proportional hazards model (Cox). We also applied time-dependent Cox (TD-Cox), which captures time-varying factors and can leverage those information to provide more accurate prediction of student dropout. Our model utilizes different groups of variables such as demographic, family background, financial, high school information, college enrollment and semester-wise credits. The proposed framework has the ability to address the challenge of predicting dropout students as well as the semester that the dropout will occur. This study enables us to perform proactive interventions in a prioritized manner where limited academic resources are available. This is critical in the student retention problem because not only correctly classifying whether a student is going to dropout is important but also when this is going to happen is crucial for a focused intervention. We evaluate our method on real student data collected at Wayne State University. Results show that the proposed Cox-based framework can predict the student dropouts and semester of dropout with high accuracy and precision compared to the other state-of-the-art methods.

Keywords

Event prediction, longitudinal data, survival analysis, student retention, classification, regression

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '16, October 24–28, 2016, Indianapolis, IN, USA.

© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...\$15.00

DOI: <http://dx.doi.org/10.1145/2983323.2983351>

1. INTRODUCTION

One of the long-term goals of any university in the U.S. and around the world is to reduce the student attrition rate [37]. It is reported that about one-fourth of the students dropped out of college after their first year and it increases to 50% by the end of the fourth semester [39]. The benefits of improving student retention is self-evident including higher chance of having a better career and higher standard of life [36]. On the other hand, the higher student retention rate, the more likely that the university is positioned higher in rankings, secure more government funds, and has easier path to program accreditations. In view of these reasons, directors and administrators in universities are increasingly feeling the pressure to outline and implement strategies to decrease student attrition. This requires a better planning for interventions and a more thorough understanding of the fundamental issues that cause the student attrition problem. In higher education, student retention rate is defined as the percentage of students who after completing a semester return to the same university for the following semester. Universities are eager to find out who are at a higher risk of dropping out from their study and how they can address this issue and improve the retention rate. Thus, this clearly motivates the need for developing predictive models that can effectively identify the students who are potentially going to dropout and the semester that the dropout is going to occur at during their college program.

Many explanatory models were found to help educational institutions to predict at-risk students [29]. Traditional methods such as regression and logistic regression have been used to identify dropout students for decades [9, 26]. Recently, student retention problem has drawn a lot of attention from researchers in data mining and machine learning communities [29, 23]. However, student attrition is not an abrupt event, but rather a lengthy process that completely depends on time [39]. Therefore, it would be appropriate to formalize it as a longitudinal problem and use sophisticated longitudinal data analysis techniques for modeling the problem. One of the important characteristics of student data is that it can be incomplete due to the inability to continuously track the student, often referred to as *censoring*. This incompleteness in events or information is different from missing data encountered in routine data mining problems and not all modeling techniques are able to handle them [41]. Ignoring the censored data on one hand yields suboptimal biased mod-

els because of neglecting available information while, on the other hand, treating censoring time as the actual event time causes underestimation of the model. Another important thing to point out is that, unlike machine learning and data mining techniques, which normally provide single outcome prediction, survival analysis estimates the survival (failure) as a function of time. In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs [25].

In spite of the success of survival analysis methods in other domains such as healthcare, engineering, etc., there is only a limited attempt of using these methods in student retention problem [32, 19]. In this paper, we propose a survival analysis based framework that uses pre-enrollment and semester-wise information to address the problem of student attrition in the presence of censored information. For this purpose, we implement *Cox* and time-dependent Cox, (*TD-Cox*) to model the student retention problem. The fundamental idea is that we can utilize the survival analysis method at an early stage of college study to predict student dropouts. Thus, the main contributions of this paper are summarized as follows:

- Rigorously define the student attrition problem and create important variables that influence this problem.
- Propose a novel student retention prediction framework to simultaneously deal with both problems, namely, “who is going to dropout” and “when the dropout will occur”.
- Using survival analysis methodology to study the temporal nature of student retention by focusing on dropout information as the outcome of interest.
- Demonstrate the performance of the proposed method using Wayne State University student data and compare with the existing state-of-the-art methods.

The rest of the paper is organized as follows. In Section 2, we describe the literature relevant to the student retention problem. After defining the notations and definitions that will be used throughout the paper, we describe the proposed Cox and TD-Cox methods for student dropout prediction in Section 3. In Section 4, we first describe the data sources that will be used in this study and then show the performance of our method on Wayne State University student data. Finally, Section 5 concludes the work and offers directions for future research.

2. RELATED WORK

Student retention is one of the most widely studied areas in higher education [39]. There are many institutions, consulting firms and businesses focusing on student retention. In the past decades, comprehensive models have been developed to address the college student attrition problem. Most of the earlier studies try to understand the reasons behind student dropout by developing theoretical models [38]. For many years, statistical methods have been used widely to predict student dropout and also to find the important factors that has an effect on the problem [44, 22]. Regression is one of the primary techniques that has been applied in this area [11]. Logistic regression is another statistical method that was frequently used in this domain [26, 9]. [27] used logistic regression, discriminant analysis and regression tree to address this issue. In another work, logistic regression

method is developed to identify freshman at risk of attrition within few weeks after freshman orientation [16]. However, these models cannot incorporate information from censored information and are likely to produce suboptimal results.

While predictive analytics has been used in other industries for many years [13], higher education is a relatively late adopter of these approaches as a tool to support making decision [40]. Recently, researchers in the area of machine learning and data mining, tried to address the student retention phenomenon [10, 42, 34]. Genetic algorithms for selecting feature subset and artificial neural networks for performance modeling have been developed to give better prediction of first year high risk students to dropout at Virginia Commonwealth University [1]. Several classification algorithms including Bayes classifier [30, 3], decision tree [18, 31, 43], boosting methods and support vector machines [45, 23] have been developed to predict student attrition with higher accuracy compared to the traditional statistical methods.

A slightly more complex relevant modeling technique is survival analysis [14]. Survival analysis is a subfield of statistics which aims at modeling longitudinal data where the outcome variable is the time until an occurrence of event [24]. In this type of regression, both components, (i) if an event (i.e. dropout) occurs or not and (ii) when the event will occur, can be incorporated [28]. Thus, the benefit of using survival analysis over logistic regression or other data mining methods is the ability to add the time component into the model and also effectively handle censored data. However, the literature in this area is limited. The use of survival analysis modeling to study both student retention and student dropout has been developed in [32, 21, 19, 20]. Among those, only [19] developed an event history model to assessing attrition behaviour among first-generation students using pre-enrollment attributes. They assume that time to dropout follows exponential distribution. However, this assumption may not be valid in many situations where time to event has more complex distribution [5].

Despite the fact that survival model have more flexibility to handle the student retention problem, there were few efforts in the literature in this student education domain. It is evident that there is considerable room for improvement in the current state-of-the-art. In this paper, we relax some of the previous assumptions including linear dropout rate of student by implementing a more rigorous survival model such as Cox proportional hazard model and also utilize time-varying features such as semester-wise GPA in more comprehensive manner by developing time-dependent Cox model. Therefore, this paper will further improve the existing ability to predict student success by showing an in-depth application of survival algorithms on student data and compare the result with other statistical and machine learning approaches which, to the best of our knowledge, has not been done anywhere else in the past.

3. PROPOSED METHOD

The primary goal of this work is to develop a time-dependent model to predict student dropout based on both pre-enrollment and semester-wise information. We also build a survival analysis framework to estimate the semester of dropout only based on pre-enrollment attributes. We begin by presenting the basic concepts and notations required to comprehend this problem. Table 1 describes the notations used in this paper.

Table 1: Notations used in this paper

Notation	Description
n	number of data points
p	number of static features
q	number of time-dependent features
X_i	$1 \times p$ matrix of features for student i
$Z_i(t)$	$1 \times q$ matrix of time-dependent features for student i
Y	$n \times 1$ vector of actual event time
C	$n \times 1$ vector of last follow-up time
T	$n \times 1$ vector of observed time which is $\min(Y, C)$
δ	$n \times 1$ binary vector of censored status
d_i	number of events occurred at time t_i
$S_0(t)$	baseline survival probability
$S(t X, Z(t))$	conditional survival probability at time t
$h_0(t)$	base hazard rate
$h(t X, Z(t))$	conditional hazard probability
β	vector of Cox regression coefficients
$L(\beta)$	maximum likelihood function for β

We will first define some of the terms that will be used in this paper.

- **Dropout Student:** It is defined as a student who does not register in a semester or whose semester GPA is zero.
- **Event:** Student dropout before his graduation is our event of interest.
- **Censored:** If student does not dropout within the first 6 semesters or by a cut-off timepoint, then it is defined as censored data.

3.1 Survival Analysis

Survival analysis is defined as a collection of statistical methods which contains time of a particular event of interest as the outcome variable to be estimated. In many survival applications, it is common to see that the observation period of interest is incomplete for some subjects and such data is considered to be *censored* [33]. Let $D_n(t) = \{X_i, Z_i(t), T_i, \delta_i(t); i = 1, \dots, n\}$ denote a sample from dataset D at time t , where X_i represents a $(1 \times p)$ covariate vector for subject i when there are p static variables in the data, $Z_i(t)$ represents $(1 \times q)$ vector of time-dependent covariates at time t and T_i denotes the observed event time. Let us suppose that Y_i is the survival time, but this may not be observed and we instead observe $T_i = \min(Y_i, C_i)$, where C_i is the censored time or the last follow-up time. We do know if the data has been censored, and together with Y_i we have the indicator variable

$$\delta_i = \begin{cases} 1 & Y_i \leq C_i \\ 0 & Y_i > C_i \end{cases}$$

So, for individual i , if $\delta_i = 0$, it is censored and if $\delta_i = 1$ it is uncensored. Figure 1 illustrates the student retention problem using survival analysis in which students A, B and D drop out before the 6th semester and students C, E and F remain at school even at the end of the 6th semester or in other words they are censored at semester 6 (shown by 'X').

Considering the duration to be a continuous random variable T , the survival function, $S(t)$ is the probability that the time of event occurs later than a certain specified time t , which is defined as

$$S(t) = \Pr(T > t) = \int_t^\infty f(u) du = 1 - F(t) \quad (1)$$

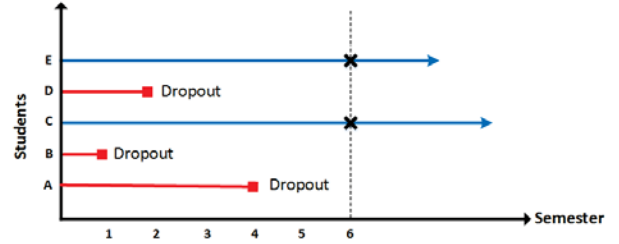


Figure 1: An illustration to demonstrate the problem of student retention. In this example, students A, B and D dropped out after 4, 1 and 2 semesters, respectively. Students C and E did not drop out in their first 6 semesters and therefore they are censored.

where $f(u)$ is a probability density function and $F(t)$ is a cumulative distribution function. An alternative characterization of the distribution of T is given by the hazard function, or instantaneous rate of occurrence of the event which is defined as

$$h(t) = \lim_{dt \rightarrow 0} \frac{\Pr(t \leq T < t + dt)}{dt} \quad (2)$$

In other words, $h(t)$ is the event rate at time t conditional on survival until time t or later. The numerator of this expression is the conditional probability that the event will occur within the interval $[t; t + dt)$ given that it has not occurred before t , and the denominator is the width of the interval. Dividing one by the other, we obtain a rate of event occurrence per unit of time. Taking the limit, as the width of the interval goes down to zero, we obtain an instantaneous rate of occurrence.

3.2 Cox Proportional Hazard Regression

One of the popular methods in survival analysis is the Cox proportional hazard model [6]. The Cox regression model is a semi-parametric technique which has fewer assumptions than typical parametric methods [4]. In particular, and in contrast with parametric models, it makes no assumptions about the shape of the baseline hazard function [8]. The Cox model provides a useful and easy way to interpret information regarding the relationship of the hazard function to predictors. The hazard function for the Cox proportional hazard model has the form

$$h(t|X) = h_0(t) \exp(\beta_1 X_1 + \dots + \beta_p X_p) = h_0(t) e^{(\beta X)} \quad (3)$$

where $h_0(t) = e^{\alpha(t)}$ is the baseline hazard function at time t and $\exp(\beta_1 X_1 + \dots + \beta_p X_p)$ is the risk associated with the covariate values. Therefore, the survival probability function for Cox model can be formulated as

$$S(t | X) = S_0(t) e^{-\beta X} \quad (4)$$

where

$$S_0(t) = e^{-\int_0^t h_0(x) dx} \quad (5)$$

Parameters of the Cox regression model are estimated by maximizing the partial likelihood [7]. Based on Cox regression formula, a partial likelihood can be constructed from the dataset as follows:

$$L(\beta) = \prod_{i: \delta_i = 1} \frac{\theta_i}{\sum_{j: t_j \geq t_i} \theta_j} \quad (6)$$

where $\theta_i = \exp(\beta X_i)$ and $(X_1, \dots, X_n)$ are the covariate vectors for the n independently sampled individuals in the dataset. By solving $\frac{\partial L(\beta)}{\partial \beta} = 0$, the covariate coefficient can be estimated as $\hat{\beta}$. To obtain the baseline hazard function, in full likelihood function, β should be replaced by $\hat{\beta}$. Thus, $h_0(t_i)$ can be obtained

$$\hat{h}_0(t_{(i)}) = \frac{1}{\sum_{j \in R(t_{(i)})} \theta_j} \quad (7)$$

3.3 Time-Dependent Cox (TD-Cox)

The Cox proportional hazard regression has an assumption that covariates are independent of time. In another words, when covariates do not change over time or when data is only collected for the covariates at one time point, it is appropriate to use static variables to explain the outcome. On the other hand, there are many situations (such as our student retention problem) where covariates change over time and the above assumption does not hold. Thus, it is more appropriate to use time-dependent covariates which result in more accurate estimates of the outcomes [15]. Consequently, we can define time-dependent variables that can change in value over the course of the observation period. Extensions to time-dependent variables can be incorporated using the counting process based formulation [2]. Essentially, in counting process, data are expanded from one record-per-student to one record-per-interval between each event time for each student. In order to have a better understanding of the counting process, we provide an illustrative example. Table 2 shows the data in record-per-student format. In this example, for each student we record the time of dropout, status and semester-wise GPA. If status is 1 it means student dropout and if it is 0, it means student did not dropout until the observed time.

Table 2: Example of survival data (GPA(1) refers to GPA for the first semester).

Student ID	Time	Status	GPA(1)	GPA(2)	GPA(3)
ID_1	1	1	2	-	-
ID_2	2	1	3.2	1.8	-
ID_3	3	0	4	4	3.5

Thus, for time-dependent survival analysis, we need to change the format using counting process. Using the first part of Algorithm 1, data is changed to record-per-interval between each event time (Table 3) per student. Basically, we consider time interval by adding t_0 column and for each interval, GPA is calculated independently. Other static variables such as demographic information which do not change over different intervals for a given student can also be appended to the same row.

Table 3: Example of survival data after counting process based reformatting.

Student ID	t_0	t	Status	GPA
ID_1	0	1	1	2
ID_2	0	1	0	3.2
ID_2	1	2	1	1.8
ID_3	0	1	0	4
ID_3	1	2	0	4
ID_3	2	3	0	3.5

In this paper, we develop Time Dependent Cox regression, namely **TD-Cox**, which can simultaneously handle both static and time-dependent covariates. Thus, the hazard function can be defined as

$$h(t|X, Z(t)) = h_0(t)e^{\beta(X+Z(t))} \quad (8)$$

Consequently, the survival probability function for TD-Cox model can be formulated as

$$S(t | X, Z(t)) = S_0(t)^{\exp(\beta(X+Z(t)))} \quad (9)$$

where $S_0(t)$ can be estimated using Eq. (5). Algorithm 1 summarizes the *TD-Cox* method. First, TD-Cox parameters are learnt using the training data based on maximum likelihood function. Then, for each student in test data, we use Eq. (9) to estimate the probability of dropout.

ALGORITHM 1: *TD-Cox method based on counting process*

Input: Student data $D_n(t) = \{X, Z(t), T, \delta\}$
Output: probability of student dropout
part 1: Reformatting Data Based on Counting Process
for $i=1$ **to** n **do**
 $T_c \leftarrow T_i$;
 for $j=0$ **to** T_c **do**
 for $k=1$ **to** q **do**
 $Z_k = Z_k(j)$;
 $t_{i+j} = j$;
 if $\delta_i=1$ **and** $j = T_c$ **then**
 $s_{i+j} = 1$;
 else
 $s_{i+j} = 0$;
 end
 end
 end
end
part 2: TD-Cox method
Learn TD-Cox parameters, β s and $\hat{h}_0$ using Eqs. (6) and (7)
for each student in the test data do
 Estimate $\hat{S}(t | X, Z)$ from Eq. (9)
 $\hat{F}(t | X, Z) = 1 - \hat{S}(t | X, Z)$
end

4. EXPERIMENTAL RESULTS

In this section, we present the results of the proposed survival analysis framework for student dropout prediction. First, we explain our data source and define the variables used in our model. We also describe the performance evaluation metrics used to compare the results of the proposed model with other methods. Finally, the results will be reported and discussed in detail.

4.1 Data Description

In this study, a dataset was compiled by tracking 11,121 students enrolled at Wayne State University (WSU) starting from 2002 until 2009. Among those, there were 16%, 18%, 8% and 11% dropouts by the end of first, second, third and fourth semesters, respectively. We only focus on FTIAC (First Time In Any College) students because the duration of study for other students (such as transferred students) is different from FTIAC students. The dependent variable is the semester of dropout. In this study, dropout is defined as a student who does not register in a semester or whose semester GPA is zero. In order to evaluate the performance of the proposed methods, we run two sets of experiments as follows:

- *Experiment 1:* In this experiment, we collected the information for students who are admitted to WSU from 2002 to 2009 and keep track of their record upto first 6

semesters. The illustration of this experiment is shown in Figure 1.

- *Experiment 2:* In this experiment, we do not follow the students for 6 full semesters. In other words, we cut the observation at Winter 2009 semester, and hence in this case, for students who have been admitted to school in 2008, we have a record for only two semesters. For better understanding, we illustrate this experiment in Figure 2. As shown in the figure, for this set of experiments, we censored the data at the 6th semester for all those students who started their degree on or before Fall 2006 semester.

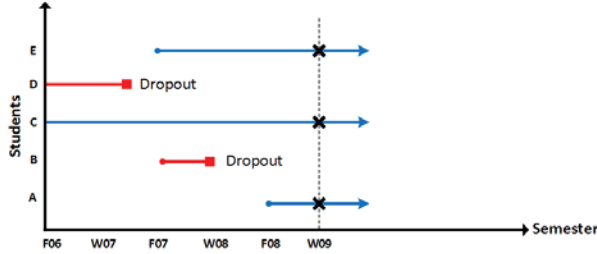


Figure 2: An illustration to demonstrate our second set of experiments. Students B and D started their degree in Fall 2007 and Fall 2006 semesters and dropped out after 1 and 2 semesters, respectively. Students A, C and E started school in Fall 2008, Fall 2006 and Fall 2007 semesters, respectively and did not drop out till 2009; so they are censored in our dataset.

After the data preparation and the necessary pre-processing, we ended up with 31 predictor attributes, which could be categorized into six different groups: demographic, family background, financial, high school information, college enrollment and semester-wise attributes (time-dependent). The complete list of attributes and their description are summarized in Table 4. For *TD-Cox* we used all 31 attributes as listed in Table 4. To have a fair comparison between *TD-Cox* and other classification methods, for each student, the average of temporal features (semester-wise attributes) before dropout is used.

Table 4: List of attributes used in this study.

List of Attributes	
Demographic Attributes:	Financial Attributes:
Gender	Student's cash amount
Marital status	Student's parents cash amount
Ethnicity	Student's income
Hispanic or non-Hispanic	Father's income
Residence county	Mother's income
Family Background Attributes:	Enrollment Attributes:
Father's education level	Age of admission
Mother's education level	First admission semester
Number of family members	Did student transfer credit?
Number of family members in college	Student's college
Pre-Enrollment Attributes:	Student's major
High School GPA	Semester-wise Attributes:
Composite ACT score	Credit hours attempts
Math ACT score	Percentage of passed credits
English ACT score	Percentage of dropped credits
Reading ACT score	Percentage of failed credits
Science ACT score	Semester GPA
High school graduation age	

4.2 Performance Evaluation

In order to have a quantitative measure for estimating the performance of the proposed model and compare with other classification techniques, we used two sets of experiments. We divide our data into training and testing sets. Training data consists of records for students who have been admitted from 2002 to 2008. The test data consists of students admitted in 2009 and they are completely unused during model building. We report the results of both 10-fold cross validation on training set and the test data in separate tables. In the first one, we use standard technique of stratified 10-fold cross validation, which divides each dataset into ten subsets, called folds, of approximately equal size and equal distribution of dropout and non-dropout students. In each experiment, one fold is used for testing the model that has been developed from the remaining nine folds during the training phase. The evaluation metrics for each method is then computed as an average of the ten experiments. We also evaluate the performance of the model learned using 2002 to 2008 data on the unseen test data, which is the dropout information for students who are admitted to WSU in 2009. We implemented our methods in R programming language using *survival* package [35] and for the rest of the methods we used open source Weka software [17]. To assess the performance of the proposed model, the following metrics are used for the classification problem:

- *Accuracy* is expressed in terms of percentage of subjects in the test set that are classified correctly.
- *F-measure* is defined as a harmonic mean of precision and recall. A high value of *F-measure* indicates that both precision and recall are reasonably high.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where $Precision = \frac{TP}{TP+FP}$, and $Recall = \frac{TP}{TP+FN}$. *TP* stands for the true positive, *FP* stands for false positive and *FN* is false negative.

- *AUC* is expressed as area under the receiver operating characteristic (ROC) where the curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) under various threshold values.

For time to dropout prediction, we applied frequently used metric in regression problem such as mean absolute error (MAE).

- *Mean Absolute Error (MAE)* is a quantity used to measure how close the predictions are to the actual outcomes. The mean absolute error is given by

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

where $\hat{y}_i$ is the predicted value and y_i is the true value for subject i .

MAE treats both underestimating and overestimating of actual value in the same manner. However, in student retention problem, these types of errors have different meaning. For example, any model that has the ability to predict the semester of dropout earlier than the actual semester has more value because, in this case, an individualized intervention programs might help to reduce the student dropout

Table 5: Performance of Logistic Regression (LR), Adaboost (AB) and Decision tree (DT) with Cox and TD-Cox on WSU student retention data from 2002 to 2008 (experiment 1) for each semester using 10-fold cross validation along with standard deviation.

	Accuracy					F-measure					AUC				
	LR	AB	DT	Cox	TD-Cox	LR	AB	DT	Cox	TD-Cox	LR	AB	DT	Cox	TD-Cox
1st Semester	0.705 (0.023)	0.709 (0.019)	0.706 (0.035)	0.719 (0.015)	0.719 (0.015)	0.702 (0.031)	0.712 (0.028)	0.700 (0.045)	0.724 (0.029)	0.724 (0.029)	0.734 (0.015)	0.747 (0.013)	0.707 (0.026)	0.751 (0.013)	0.705 (0.013)
2nd Semester	0.731 (0.025)	0.737 (0.023)	0.725 (0.037)	0.742 (0.019)	0.755 (0.018)	0.726 (0.033)	0.729 (0.029)	0.719 (0.051)	0.738 (0.031)	0.741 (0.03)	0.772 (0.018)	0.783 (0.015)	0.724 (0.023)	0.786 (0.013)	0.792 (0.011)
3rd Semester	0.751 (0.024)	0.757 (0.019)	0.734 (0.034)	0.769 (0.018)	0.778 (0.016)	0.749 (0.034)	0.746 (0.033)	0.734 (0.048)	0.754 (0.028)	0.761 (0.029)	0.803 (0.017)	0.805 (0.016)	0.737 (0.019)	0.808 (0.013)	0.811 (0.012)
4th Semester	0.772 (0.025)	0.768 (0.027)	0.741 (0.031)	0.781 (0.017)	0.801 (0.018)	0.764 (0.036)	0.759 (0.032)	0.739 (0.038)	0.773 (0.029)	0.784 (0.027)	0.821 (0.016)	0.827 (0.018)	0.758 (0.021)	0.832 (0.014)	0.835 (0.013)
5th Semester	0.783 (0.025)	0.775 (0.029)	0.753 (0.039)	0.796 (0.02)	0.812 (0.019)	0.779 (0.034)	0.771 (0.031)	0.749 (0.039)	0.784 (0.027)	0.803 (0.026)	0.827 (0.016)	0.829 (0.015)	0.769 (0.019)	0.835 (0.012)	0.84 (0.012)
6th Semester	0.798 (0.028)	0.789 (0.024)	0.775 (0.031)	0.803 (0.017)	0.821 (0.015)	0.796 (0.03)	0.785 (0.035)	0.769 (0.043)	0.800 (0.028)	0.818 (0.024)	0.837 (0.012)	0.835 (0.015)	0.787 (0.021)	0.840 (0.011)	0.847 (0.009)

Table 6: Performance of Logistic Regression (LR), Adaboost (AB) and Decision tree (DT) with Cox and TD-Cox on 2009 WSU student retention (experiment 1) for each semester along with standard deviation values.

	Accuracy					F-measure					AUC				
	LR	AB	DT	Cox	TD-Cox	LR	AB	DT	Cox	TD-Cox	LR	AB	DT	Cox	TD-Cox
1st Semester	0.701 (0.019)	0.709 (0.017)	0.689 (0.025)	0.715 (0.018)	0.715 (0.018)	0.703 (0.025)	0.710 (0.027)	0.697 (0.03)	0.719 (0.024)	0.719 (0.024)	0.716 (0.016)	0.723 (0.012)	0.701 (0.017)	0.742 (0.014)	0.742 (0.014)
2nd Semester	0.727 (0.018)	0.731 (0.019)	0.717 (0.025)	0.739 (0.019)	0.745 (0.02)	0.721 (0.029)	0.723 (0.026)	0.711 (0.029)	0.732 (0.021)	0.745 (0.023)	0.753 (0.015)	0.769 (0.013)	0.725 (0.019)	0.783 (0.015)	0.787 (0.016)
3rd Semester	0.747 (0.019)	0.745 (0.022)	0.729 (0.024)	0.757 (0.015)	0.773 (0.016)	0.747 (0.027)	0.740 (0.025)	0.725 (0.032)	0.751 (0.023)	0.768 (0.021)	0.779 (0.014)	0.785 (0.013)	0.729 (0.018)	0.797 (0.013)	0.805 (0.012)
4th Semester	0.768 (0.021)	0.761 (0.018)	0.737 (0.023)	0.778 (0.017)	0.784 (0.014)	0.759 (0.026)	0.754 (0.024)	0.732 (0.031)	0.765 (0.023)	0.789 (0.024)	0.817 (0.014)	0.819 (0.016)	0.748 (0.017)	0.824 (0.011)	0.828 (0.014)
5th Semester	0.779 (0.017)	0.770 (0.019)	0.748 (0.021)	0.785 (0.017)	0.805 (0.016)	0.771 (0.023)	0.766 (0.028)	0.739 (0.031)	0.789 (0.025)	0.805 (0.022)	0.824 (0.012)	0.819 (0.014)	0.76 (0.018)	0.83 (0.012)	0.839 (0.012)
6th Semester	0.791 (0.018)	0.778 (0.016)	0.769 (0.02)	0.799 (0.014)	0.83 (0.013)	0.788 (0.024)	0.779 (0.027)	0.761 (0.029)	0.796 (0.022)	0.814 (0.02)	0.832 (0.014)	0.829 (0.015)	0.773 (0.017)	0.836 (0.010)	0.844 (0.008)

rate. Therefore, we also evaluated the models using the following two domain based metrics:

- *Underestimated Prediction Error Rate (UPER)* is defined as the fraction of the underestimated prediction output over the entire prediction error.

$$UPER = \frac{\sum_{i=1}^n I(\hat{y}_i < y_i)}{\sum_{i=1}^n I(\hat{y}_i < y_i) + \sum_{i=1}^n I(\hat{y}_i > y_i)}$$

- *Overestimated Prediction Error Rate (OPER)*: since the total number of error is a constant, OPER can be calculated as

$$OPER = 1 - UPER$$

It should be noted that, in the student retention problem, any model with higher UPER than OPER will be of great interest because it is better to underestimate the semester of dropout earlier, rather than overestimate it.

4.3 Results and Discussion

We show the experimental result for two types of analyses: “predicting dropout student” and “estimating semester of dropout”. Also the practical implications of our framework in educational studies will be discussed in this section.

4.3.1 Predicting Dropout Students

We compare the performance of our proposed TD-Cox and the standard Cox method against three well-known classification techniques in the machine learning domain: Logistic Regression (LR), Adaptive Boosting (AB) and Decision Tree (DT). We test the performance of the models to predict the student dropout in different semesters for the two experimental setups explained in Section 4.1. The results are shown in Tables 5-8. From these Tables, we can see the best performance and consistent result of the *TD-Cox* method. In this study, as described in Table 4, we define 5 after-enrollment variables including GPA, percentage of passed, dropped or failed credits and credit hours attempts. When we used those attributes along with pre-enrollment variables in the proposed *TD-Cox* method, we get better classification performance. Thus, unlike other classification methods, the proposed *TD-Cox* approach has the ability of using extra semester-wise information by introducing time-dependent variables in the model.

Figures 3 and 4 provide the performance comparison between all the methods using each semester for different experimental setups. It can be observed that, the accuracy and F-measure increase significantly for *TD-Cox* when we have more semester-wise information. The ability of *TD-Cox* to leverage those information provides more accurate prediction of student dropout. We can also conclude that in the presence of censored data, survival analysis methods such as the one that is being used in this paper (*Cox* and *TD-Cox*) are a better choice in order to predict student

Table 7: Performance of Logistic Regression (LR), Adaboost (AB) and Decision tree (DT) with Cox and TD-Cox on WSU student retention data from 2002 to 2008 (experiment 2) for each semester using 10-fold cross validation along with standard deviation.

	Accuracy					F-measure					AUC				
	LR	AB	DT	Cox	TD-Cox	LR	AB	DT	Cox	TD-Cox	LR	AB	DT	Cox	TD-Cox
1st Semester	0.705 (0.023)	0.709 (0.019)	0.706 (0.035)	0.719 (0.015)	0.719 (0.015)	0.702 (0.031)	0.712 (0.028)	0.7 (0.049)	0.724 (0.029)	0.724 (0.029)	0.734 (0.015)	0.747 (0.013)	0.707 (0.026)	0.751 (0.013)	0.751 (0.013)
2nd Semester	0.731 (0.025)	0.737 (0.026)	0.725 (0.037)	0.742 (0.019)	0.751 (0.018)	0.726 (0.033)	0.729 (0.029)	0.719 (0.051)	0.738 (0.031)	0.741 (0.03)	0.772 (0.018)	0.783 (0.015)	0.724 (0.023)	0.786 (0.013)	0.792 (0.011)
3rd Semester	0.739 (0.021)	0.743 (0.021)	0.729 (0.036)	0.757 (0.019)	0.768 (0.017)	0.734 (0.036)	0.74 (0.033)	0.723 (0.043)	0.749 (0.029)	0.752 (0.029)	0.789 (0.016)	0.798 (0.014)	0.731 (0.02)	0.805 (0.018)	0.816 (0.014)
4th Semester	0.758 (0.023)	0.753 (0.027)	0.731 (0.034)	0.767 (0.018)	0.777 (0.017)	0.750 (0.033)	0.751 (0.035)	0.729 (0.048)	0.765 (0.028)	0.774 (0.029)	0.811 (0.018)	0.814 (0.016)	0.745 (0.023)	0.818 (0.016)	0.828 (0.013)
5th Semester	0.723 (0.027)	0.762 (0.03)	0.739 (0.039)	0.785 (0.021)	0.796 (0.019)	0.762 (0.035)	0.759 (0.032)	0.735 (0.041)	0.779 (0.027)	0.79 (0.027)	0.817 (0.016)	0.819 (0.017)	0.758 (0.018)	0.825 (0.015)	0.838 (0.013)
6th Semester	0.781 (0.031)	0.77 (0.025)	0.753 (0.035)	0.792 (0.018)	0.815 (0.017)	0.773 (0.033)	0.766 (0.035)	0.743 (0.046)	0.787 (0.026)	0.811 (0.025)	0.827 (0.015)	0.83 (0.015)	0.769 (0.022)	0.831 (0.013)	0.84 (0.009)

Table 8: Performance of Logistic Regression (LR), Adaboost (AB) and Decision tree (DT) with Cox and TD-Cox on 2009 WSU student retention (experiment 2) for each semester along with standard deviation.

	Accuracy					F-measure					AUC				
	LR	AB	DT	Cox	TD-Cox	LR	AB	DT	Cox	TD-Cox	LR	AB	DT	Cox	TD-Cox
1st Semester	0.701 (0.019)	0.709 (0.017)	0.689 (0.025)	0.715 (0.018)	0.715 (0.018)	0.703 (0.025)	0.71 (0.027)	0.697 (0.03)	0.719 (0.024)	0.719 (0.024)	0.716 (0.016)	0.723 (0.012)	0.701 (0.017)	0.742 (0.014)	0.742 (0.014)
2nd Semester	0.727 (0.018)	0.731 (0.019)	0.717 (0.022)	0.739 (0.019)	0.745 (0.02)	0.721 (0.029)	0.723 (0.026)	0.711 (0.029)	0.732 (0.021)	0.745 (0.023)	0.743 (0.015)	0.759 (0.013)	0.725 (0.019)	0.773 (0.017)	0.787 (0.016)
3rd Semester	0.735 (0.022)	0.740 (0.024)	0.725 (0.026)	0.753 (0.018)	0.767 (0.017)	0.739 (0.029)	0.734 (0.023)	0.725 (0.035)	0.749 (0.023)	0.768 (0.021)	0.768 (0.015)	0.772 (0.015)	0.728 (0.018)	0.789 (0.014)	0.801 (0.013)
4th Semester	0.752 (0.023)	0.745 (0.019)	0.728 (0.024)	0.764 (0.02)	0.774 (0.016)	0.747 (0.031)	0.743 (0.025)	0.731 (0.035)	0.768 (0.025)	0.789 (0.024)	0.793 (0.018)	0.798 (0.018)	0.735 (0.019)	0.816 (0.015)	0.820 (0.016)
5th Semester	0.768 (0.019)	0.762 (0.018)	0.739 (0.023)	0.781 (0.018)	0.792 (0.017)	0.765 (0.025)	0.759 (0.029)	0.742 (0.031)	0.779 (0.029)	0.809 (0.022)	0.813 (0.015)	0.805 (0.016)	0.749 (0.021)	0.827 (0.013)	0.835 (0.012)
6th Semester	0.784 (0.017)	0.769 (0.019)	0.748 (0.021)	0.797 (0.015)	0.820 (0.013)	0.778 (0.027)	0.773 (0.028)	0.754 (0.029)	0.787 (0.024)	0.818 (0.021)	0.825 (0.013)	0.819 (0.017)	0.758 (0.019)	0.83 (0.01)	0.839 (0.009)

dropout. On the other hand, even if we rely only on the pre-enrollment attributes, *Cox* provides a better performance compared to other classification methods. From the results, the Cox based methods improve the prediction accuracy of student dropout by approximately 9%. This suggests that Cox regression model would be a better choice for longitudinal data classification problems compared to the traditional methods. One important reason behind this is that it can appropriately handle censored data. Thus, it is important to note that time-dependent variables and tackling censored data are two specific features of this student retention data that survival models such as Cox can efficiently handle.

4.3.2 Estimating Semester of Dropout

One of the primary objectives of this work, is to build a model to estimate the semester of dropout at the beginning of the study. As discussed earlier, one of the drawbacks of using linear regression in the presence of censored data is that this information cannot be handled properly thus resulting in a biased estimation of time to dropout for the student retention problem. Therefore, the standard classification and regression methods will not be able to answer the important question of “when a student is going to dropout?” in the presence of censored data. In this paper, we apply our survival analysis based framework to answer this question. Table 9 shows the result of 10-fold cross-validation training data and 2009 data as test data, using first and second experimental setups. We compare the result of *Cox* with linear regression and well-known Support Vector Regression

(SVR) [12]. We should also mention that *TD-Cox* cannot be used for this purpose as we only want to use pre-enrollment information to estimate the semester of dropout. *TD-Cox* uses semester-wise information which are available only after the students begin their semester. In other words, we are interested in estimating the semester of dropout without using any semester-wise information (After-enrollment variables).

From our results, we can conclude that the *Cox* model outperforms other methods. From Table 9, it is clear that, in the presence of censored data, survival based methods such as *Cox* model have a better performance compared to the traditional regression methods. We can also observe that *Cox* has the higher UPER, which indicates that majority of errors come from underestimating the semester of dropout. This will allow us to have a better individualized intervention programs with more focus towards specific high-risk students as early as (s)he starts the school. Consequently, we are able to maximize the retention rate which can then translate into increasing number of graduations from the university.

4.3.3 Practical Implications of Our Framework

University administrators can deploy the results of proposed methods to predict students at the risk of dropout early in their academic career. Our study has shown the benefits of survival analysis as a methodology for the study of college student dropout behaviors. As mentioned earlier,

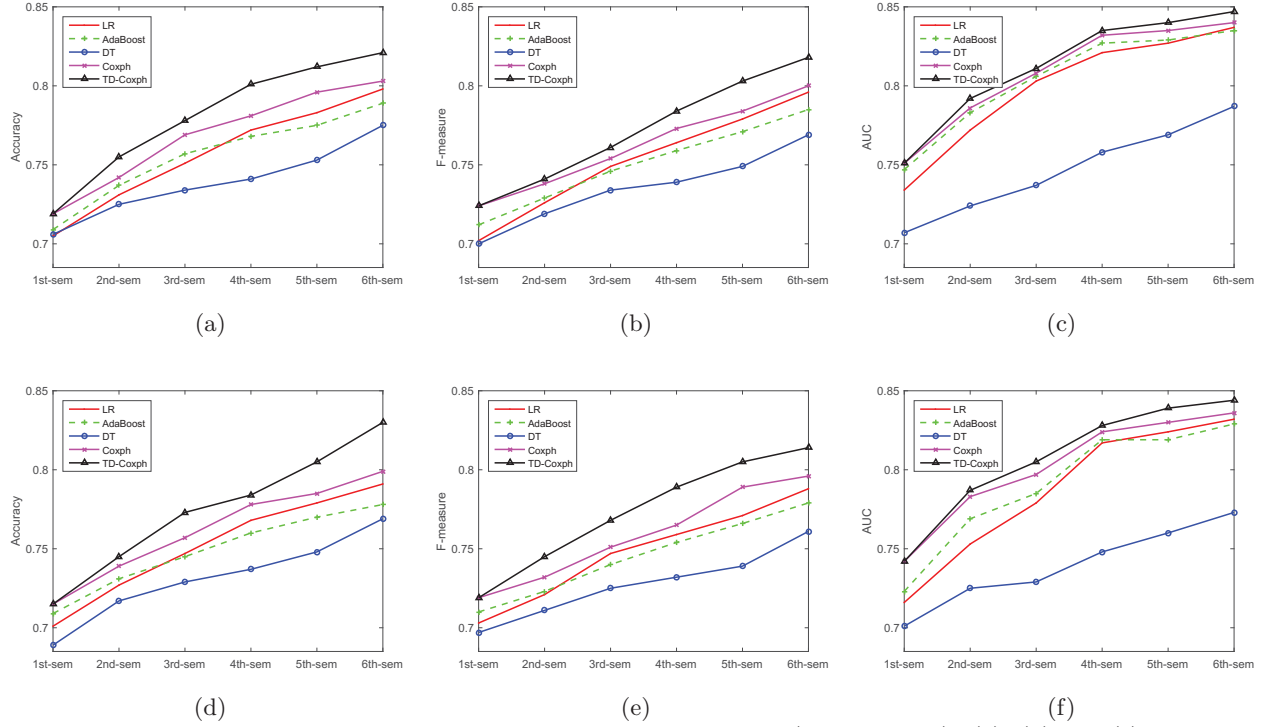


Figure 3: Performance of different methods obtained at different semesters (experiment 1): (a), (b) and (c) are the results of 10-fold cross validation on 2002-2008 training data based on different evaluation metrics and (d), (e) and (f) are the corresponding results for the 2009 test dataset.

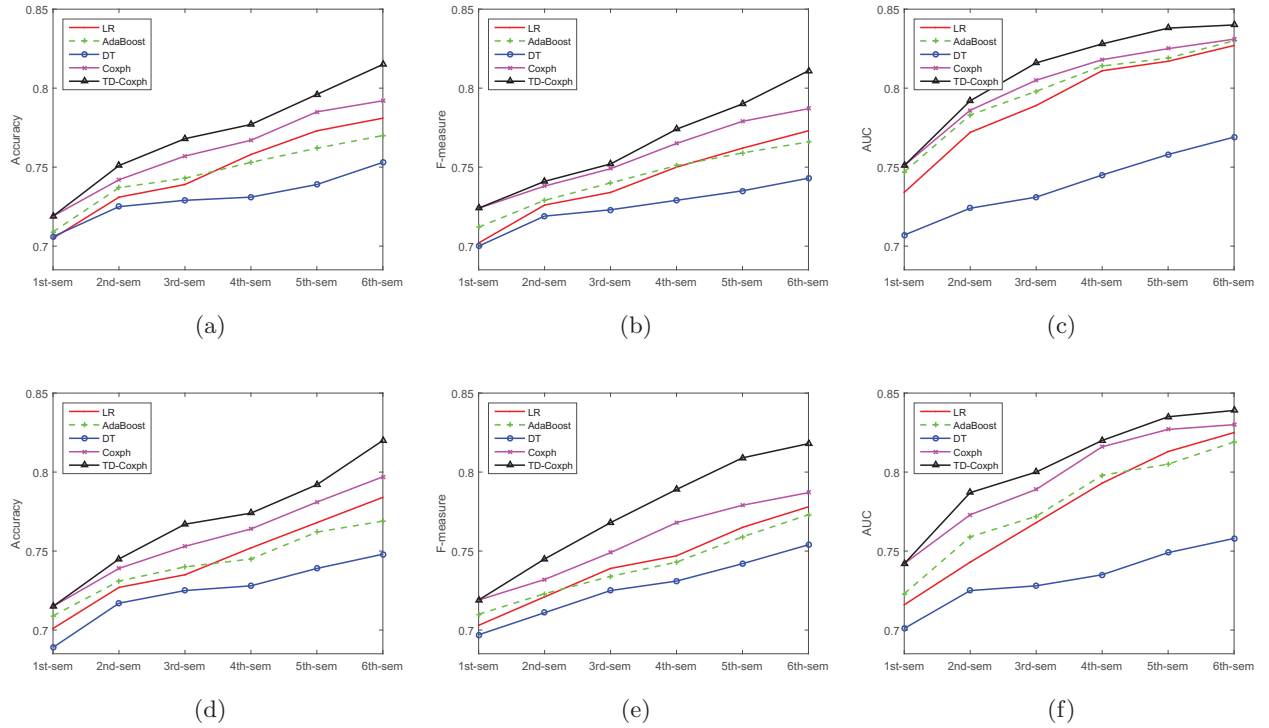


Figure 4: Performance of different methods obtained at different semesters (experiment 2): (a), (b) and (c) are the results of 10-fold cross validation on 2002-2008 training data based on different evaluation metrics and (d), (e) and (f) are the corresponding results for the 2009 test dataset.

Table 9: Performance of linear regression, SVR and Cox methods in predicting the semester of dropout on WSU student retention data from 2002 to 2008 for each semester using 10-fold cross validation and 2009 student retention data based on both experiments (exp1 and exp2).

		10-fold Cross Validation			Test Data (Year 2009)		
		MAE	UPER	OPER	MAE	UPER	OPER
Exp 1	Regression	1.79	0.623	0.377	1.83	0.612	0.388
	SVR	1.83	0.582	0.418	1.92	0.575	0.4225
	Cox	1.07	0.679	0.321	1.09	0.668	0.332
Exp 2	Regression	1.91	0.608	0.392	1.98	0.603	0.397
	SVR	1.97	0.579	0.421	2.04	0.572	0.428
	Cox	1.12	0.657	0.343	1.13	0.652	0.348

by applying Cox based methods, we are able to improve the accuracy of predicting dropout student by 9% which, in our dataset, means predicting approximately 500 at-risk students more compared to the state-of-the-art methods. Interactions with educators revealed that the ability to build a model that provides the prediction result at an early stage with high accuracy and also ranked the important factors that cause student to dropout is very crucial. Therefore, in this paper, we incorporated both pre-enrollment information and the semester-wise data to develop a survival analysis framework to be able to predict students who are going to dropout and the semester of dropout in their early college life. This can help colleges and universities to design effective retention strategies in order to compel students persisting toward degree completion. For instance, if the result of applying *Cox*-based methods show that student A is at risk due to academic reasons such as high school GPA and student B is at risk due to financial reasons, then retention support program for student A would be different than B. On the other hand, if the model predicts that student C has a higher chance of dropout by second semester while the student D has a lower chance of dropping out by the first four semesters, university administrators should give the student C a higher priority compared to student D for addressing the issue. Hence, our work will enable universities to utilize their resources more efficiently by targeting only the high risk students who are more vulnerable of dropping out of their study at any given semester. The findings of our current study support the intuitively appealing conclusion that background academic strengths such as high school GPA, semester-wise information, such as GPA and percentage of failed credit and financial attributes, have the highest impact on student to drop out from school.

5. CONCLUSION

Predicting students who will dropout from their study is an important and challenging task for academic institutions. However, little research in higher education has focused on using the data mining and statistical methods for predicting student retention. College student attrition is a longitudinal process, which enforces the need for a longitudinal modeling approach. Benefits of survival analysis as an approach for estimating the time of critical events are clear in many different application domains. In this paper, we develop a survival analysis based framework for the problem of estimating the students who are at high risk of dropping out from their study at early stage of their higher education. In our work, extending survival analysis to the study of reten-

tion has provided the ability to study the temporal nature of the attrition behaviors.

Predictive analytics holds significant promise for helping higher education institutes to make evidence-based decisions about student life cycle such as dropout problem. Motivated by this work at Wayne State University, the proposed method allows educational institutions to undertake timely measures and actions on their student attrition challenge. Once identified, these at-risk students can then be targeted with academic and administrative support to increase their chance of staying in the program. Based on the findings of this study, we can use pre-enrollment information such as screening test to identify students who are at a higher risk of dropping out of their study. It also shows that one can use the number of withdrawn or passed credits and GPA at each semester as an early warning to effectively intervene when the students are doing poorly. The survival analysis methods applied in this work are focused on early student dropout phenomenon. However, the proposed framework can be extended to model late student dropout problem using other factors that have an influence on the risk of senior students dropout. It is recommended that future research on student retention behaviors should be conducted using other available information such as course interaction websites which contain student activity information for each course. This can help with developing better interventions that can be deployed early on during a course to improve student success which can drastically reduce the student dropouts.

Acknowledgments

This work was supported in part by the US National Science Foundation grants IIS-1231742, IIS-1527827 and IIS-1646881.

6. REFERENCES

- [1] R. Alkhasawneh. *Developing a hybrid model to predict student first year retention and academic success in STEM disciplines using neural networks*. PhD thesis, Virginia Commonwealth University, 2011.
- [2] P. Andersen and R. Gill. Cox regression model for counting process: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.
- [3] B. K. Bhardwaj and S. Pal. Data mining: A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security*, 9(4):136–140, 2011.
- [4] N. E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57, 1975.
- [5] K. J. Carroll. On the use and utility of the weibull model in the analysis of survival data. *Controlled clinical trials*, 24(6):682–701, 2003.
- [6] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.
- [7] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.
- [8] D. R. Cox. Some remarks on the analysis of survival data. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 1–9. Springer, 1997.
- [9] M. S. DeBerard, G. Spielmans, and D. Julka. Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College student journal*, 38(1):66–80, 2004.
- [10] D. Delen. Predicting student attrition with data

- mining methods. *Journal of College Student Retention: Research, Theory & Practice*, 13(1):17–35, 2011.
- [11] E. L. Dey and A. W. Astin. Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, 34(5):569–581, 1993.
- [12] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems*, pages 155–161, 1997.
- [13] M. J. Fard, S. Ameri, and A. Zeinal Hamadani. Bayesian approach for early stage reliability prediction of evolutionary products. In *Proceedings of the International Conference on Operations Excellence and Service Engineering*, pages 361–371. Orlando, Florida, USA, 2015.
- [14] M. J. Fard, S. Chawla, and C. K. Reddy. Early-stage event prediction for longitudinal data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 139–151. Springer, 2016.
- [15] L. D. Fisher and D. Y. Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157, 1999.
- [16] J. G. Glynn, P. L. Sauer, and T. E. Miller. A logistic regression model for the enhancement of student retention: The identification of at-risk freshmen. *International Business & Economics Research Journal (IBER)*, 1(8), 2011.
- [17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.
- [18] S. Herzog. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research*, 2006(131):17–33, 2006.
- [19] T. T. Ishitani. A longitudinal approach to assessing attrition behavior among first-generation students: Time-varying effects of pre-college characteristics. *Research in higher education*, 44(4):433–449, 2003.
- [20] T. T. Ishitani. Studying attrition and degree completion behavior among first-generation college students in the united states. *Journal of Higher Education*, 77(5):861–885, 2006.
- [21] T. T. Ishitani and S. L. DesJardins. A longitudinal investigation of dropout from college in the united states. *Journal of college student retention: research, theory & Practice*, 4(2):173–201, 2002.
- [22] D. R. Jones-White, P. M. Radcliffe, R. L. Huesman Jr, and J. P. Kellogg. Redefining student success: Applying different multinomial regression techniques for the study of student graduation across institutions of higher education. *Research in Higher Education*, 51(2):154–174, 2010.
- [23] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1909–1918. ACM, 2015.
- [24] E. T. Lee and J. Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003.
- [25] Y. Li, J. Wang, J. Ye, and C. K. Reddy. A multi-task learning formulation for survival analysis. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1715–1724, 2016.
- [26] J. Lin, P. Imbrie, and K. J. Reid. Student retention modelling: An evaluation of different methods and their impact on prediction results. *Research in Engineering Education Symposium*, pages 1–6, 2009.
- [27] J. Luna. *Predicting student retention and academic success at New Mexico Tech*. PhD thesis, New Mexico Institute of Mining and Technology, 2000.
- [28] P. A. Murtaugh, L. D. Burns, and J. Schuster. Predicting the retention of university students. *Research in higher education*, 40(3):355–371, 1999.
- [29] A. Nandeshwar, T. Menzies, and A. Nelson. Learning patterns of university student retention. *Expert Systems with Applications*, 38(12):14984–14996, 2011.
- [30] U. K. Pandey and S. Pal. Data mining: A prediction of performer or underperformer using classification. *International Journal of Computer Science and Information Technology*, 2(2):686–690, 2011.
- [31] M. Quadri and N. Kalyankar. Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, 10(2):1–5, 2010.
- [32] P. Radcliffe, R. Huesman, J. Kellogg, and D. Jones-White. Identifying students at risk: Utilizing survival analysis to study student athlete attrition. In *National Symposium on Student Retention, Albuquerque, NM*, pages 1–11, 2006.
- [33] C. K. Reddy and Y. Li. A review of clinical prediction models. In C. K. Reddy and C. C. Aggarwal, editors, *Healthcare Data Analytics*. Chapman and Hall/CRC Press, 2015.
- [34] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2):321–330, 2014.
- [35] T. M. Therneau. *A Package for Survival Analysis in S*, 2015. version 2.38.
- [36] L. Thomas. Student retention in higher education: the role of institutional habitus. *Journal of Education Policy*, 17(4):423–442, 2002.
- [37] V. Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125, 1975.
- [38] V. Tinto. *Leaving college: Rethinking the causes and cures of student attrition*. ERIC, 1987.
- [39] V. Tinto. Research and practice of student retention: what next? *Journal of College Student Retention: Research, Theory & Practice*, 8(1):1–19, 2006.
- [40] A. Van Barneveld, K. E. Arnold, and J. P. Campbell. Analytics in higher education: Establishing a common language. *EDUCAUSE learning initiative*, 1:1–11, 2012.
- [41] B. Vinzamuri and C. K. Reddy. Cox regression with correlation based regularization for electronic health records. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 757–766, 2013.
- [42] S. K. Yadav, B. Bharadwaj, and S. Pal. Mining education data to predict student's retention: A comparative study. *International Journal of Computer Science and Information Security*, 10(2):113, 2012.
- [43] C. H. Yu, S. DiGangi, A. Jannasch-Pennell, and C. Kaprolet. A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8:307–325, 2010.
- [44] G. Zhang, T. J. Anderson, M. W. Ohland, and B. R. Thorndyke. Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. *Journal of Engineering education*, 93(4):313–320, 2004.
- [45] Y. Zhang, S. Oussena, T. Clark, and K. Hyensook. Using data mining to improve student retention in higher education: a case study. In *12th International Conference on Enterprise Information Systems*, pages 1–8, 2010.