DIGITALCOMMONS
@WAYNESTATE

Wayne State University

Wayne State University Theses

1-1-2015

Survival Analysis Approach For Early Prediction Of Student Dropout

Sattar Ameri
Wayne State University,

SURVIVAL ANALYSIS APPROACH FOR EARLY PREDICTION OF STUDENT DROPOUT

by

SATTAR AMERI

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2015

MAJOR: COMPUTER SCIENCE

Approved by:

____

Advisor Date

DEDICATION

To my dear wife for her devoted support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

One of the long-term goals of any university in the U.S. and around the world is increasing the student retention. The negative impacts of student dropout are clear to students, parents, universities and society. The positive effect of decreasing student attrition is also self-evident including higher chance of having a better career and higher standard of life for college graduates. Not only from student perspective but also college rankings, federal funding agencies and state appropriation committees are all directly dependent on by student retention rates. Thus, the higher the student retention rate, the more likely that the university is positioned higher in the ranking, secure more government funds, and have easier path to program accreditations. In view of these reasons, directors in higher education feel increasingly pressurized to outline and implement strategies to increase student retention.

In this thesis, firstly, we provide a detailed analysis of the student attrition problem and predict the risk of dropout at Wayne State University. Methods that are currently being used for the problem of student dropout are the standard preliminary statistical approaches. Our work has a number of advantages with the potential of being employed by higher education administrators at various universities. We take advantage of multiple kinds of information about different aspects of student's characteristic and efficiently utilize them to make a personalized decision about the risk of dropout for a particular student.

In the second part of this thesis, we propose survival analysis model for the student retention problem. Survival analysis method has been shown to be successful in other applications such as healthcare and biostatistics. Although well suited as a statistical technique to study student dropout, survival analysis has not been used efficiently in this study domain. With survival analysis, we also address the challenge of “time to event occurrence”. This is critical in the student retention problems because not only correctly classifying whether student is going to dropout or not is important but also when this is going to happen is crucial to investigate for further interventions. In such cases, the reliable estimation of risk at early stage of student education is very important. We propose a novel framework that uses both pre-enrollment and semester-wise information to address this issue. The basic idea here is to utilize the survival analysis method at early stage of college study to predict student success.

1.2 Motivation

The problem of college student retention is very important for educational researchers, managers, and the higher education members. Costs accrue to society, the institution, and the student when the student degree completion is not realized [30]. From a societal perspective, the achievement of a college diploma improves mediate background resources, like family economic situation. Thus, it impacts the subsequent occupational status, potential earnings, and social status achievement [37]. This fact is proved by the difference between social situation achievements of every single person from the same level of economic condition with different levels of educational degree [30]. From the institution’s perspective, maintaining enrollments is important for the economic stability. As the number of new students have fluctuated, finding the characteristics of students that remain enrolled and graduate is very important. Figure 1.1 demonstrates a process flowchart of graduation for a college student. Universities and colleges and all other institutions become fully aware of this fact that the primary assets needed to recruit, enroll, enlist, register, advise and assist a new student are the same whether that student remains and graduates or not. For instance, publication and marketing expenses, costs associated with maintaining a staff of professional counselors, travel costs connected with conducting college fairs and informational meetings, costs related with tele-counseling, staff time contacting students for yield enhancement, and staff time conducting academic advising, are all examples of pre-enrollment expenses experienced by the institution. Altogether, finding a better understanding of which college students are more likely to register and remain enrolled, is significant for maximizing those pre-enrollment resources and generating revenue.

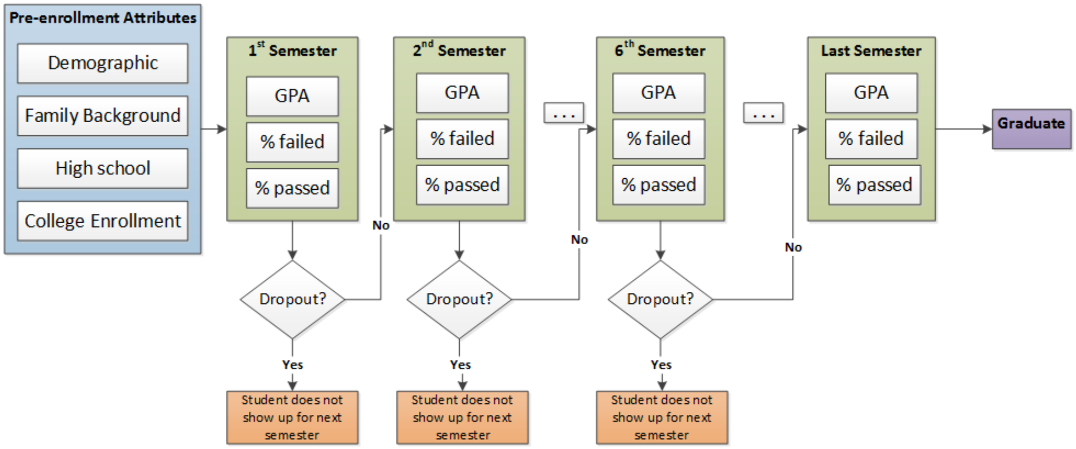


Figure 1.1: Process flowchart of graduation for a college student.

Understanding student graduation behaviours can influence the improvement or enhancement of retention strategies and prompting a higher graduation rate. In addition, graduation rates are progressively employed as a measure of a college's efficiency and are sometimes linked to resource allocation. Graduation data is needed as a component of the “Common Data Set”, a reliable set of information given by individual institutions as a way for future college students and their families to make reasonable comparison among other institutions. Graduation information is also needed by other media, like “World Report” or “US News” to give rating to institutions and provide an extra asset for comparison. Thirty states including Michigan have a funding formula that allocates some amount of funding according to efficiency indicators like time to degree, transfer rates, the number of minority and low-income graduates, the number of degrees awarded and course completion. More precisely this formula considers 6 parameters for total score for funding:

1- Undergraduate degree completions in critical skill areas

2- Research and development expenditures

3- Six-year graduation rate

4- Total degree completions

5- Institutional support expenditures as a percentage of total core expenditures

6- Pell Grant students where in Fiscal year 2014-15 in Michigan 37.3 million dollars for universities and 8.9 million dollar for community colleges were allocated based on the above performance metrics. [1]

In higher education we define student retention rate as the percentage of students who after completing a semester, return to the same university for the following semester. Student retention rate not only has effect on the university but also can affect the nearby cites. An institution's retention rate influences the job opportunities for students and public opinion. Universities are eager to find out which factors or attributes are important for their students' retention, how they can address this issue and with help of this kind of analysis they can improve their retention rate. It is important because higher retention rate will increase university funding from state and also help them to recruit more students and faculty in order to give better service to the community. College student graduation rates are often used as a measure of institution's performance. However, on the other hand, the dropout rate can show university failure to persuade student successfully finish their school. These kinds of analyses play a significant role for universities' decision about how to expand their majors' capacities and how to allocate financial aid between students. It is not beneficial to give financial aid to who are at high risk of drop out. More important is that the universities are accountable for attrition rate in their

[1] www.house.mi.gov/hfa

colleges and hence they will be penalized with narrowing their funds. In recent years, the cost of higher education has increased continuously and the amount of fund from state and federal governments that colleges received is decreased and hence the universities should try to spend their funds very wisely and doing this without analytical studies is not feasible.

From the university point of view, maintaining enrollment is important for economic stability, institutional success and managing resource allocation. From the students' perspective, it will result in higher chance of graduation that leads to higher chance of getting a good job, earn more money and a better life style in the future. In general, knowing the reasons for student dropout can help the faculty and administrators to take necessary actions so that the success percentage can be improved.

1.3 Problem Statement

First-year students have significantly increased during recent years and consequently it resulted in a huge volume of educational data. Thousands of students are admitted to study at universities every year but after first year of study some of them dropout from school. Thus, monitoring and supporting them is a topic that needs consideration at many educational institutions. University staff would like to encourage such students to finish their studies but it is hard to identify them at their early stages. It is important to explore effective approaches for predicting student dropout as well as identifying the factors affecting it with a sufficiently high accuracy. As an example at the College of Engineering of Wayne State University, for the academic year 2012, the dropout rate of freshmen is about 25% in the first year and it increases to 35% after passing two years of study which shows the importance of modeling student dropout early during their study. This thesis is intended to not only find out whether a student drop out or not but also aims at estimating the semester of dropout using survival analysis model.

There is not one single way to define student success in higher education, however, the most common measure in the academic research domain is retention rate. Students may stay in school and graduate or they do not. Thus, retention models usually predict a binary dependent variable, whether students dropout (coded as “1”) or do not dropout (coded as “0”). Analysts then typically build models using any predictive methods that are appropriate when the dependent variable is binary. Thus, in this thesis we try to answer two questions; first, which students are going to dropout? (using both pre-enrollment information and semester-wise data) and second, when the student is going to dropout? For the first question we develop survival analysis method: Cox and TD-Cox (time dependent Cox) and compare the result with other classification methods such as decision tree, adaptive boosting, and logistic regression and for the second question, we develop Cox model and compare the result with linear regression and support vector regression where censored data are not considered. However, one question that arises is “Why previous classification and regression methods are not appropriate to use for the student retention problem”. Basically, as mentioned earlier, survival analysis methods were shown to be successful in other applications such as healthcare and biostatistics but has been used very infrequently for this research problem. In the presence of censored data, the traditional methods such as linear regression or logistic regression typically fail because these methods cannot consider observations with censored data. Hence, in this work, a novel framework is proposed to use both pre-enrollment and semester-wise information to address the student dropout problem issue.

1.4 Thesis Contributions

To find an answer for the student retention problem discussed above, we introduce an intuitive survival analysis technique. Thus, the main contributions of this thesis are summarized as follows:

- Rigorously define the student attrition problem and create important variables that influence the student dropout.

- Propose a novel early student retention framework which can deal with both questions: “who is going to dropout” and “when the dropout occurs”.

- Using survival analysis methodology to study the temporal nature of student retention by incorporating semester-wise student information into the model and focusing on dropout information as the outcome of interest.

- Demonstrate the performance of the proposed method using Wayne State University student enrollment data and compare with the existing state-of-the-art methods.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows. In chapter 2 we define student dropout problem and explain the important variables along with some standard prediction methods. In chapter 3, we propose a student dropout prediction model based on survival analysis methods. Chapter 4 demonstrates the experimental results and shows the practical significance of our work using Wayne State University student data. Finally, chapter 5 concludes our discussion along with some future research directions in this area.

CHAPTER 2

STUDENT DROPOUT PREDICTION

In higher education, student dropout rate can be defined as the percentage of students who, after completing a semester, do not return to the same university for the following semester. Universities are eager to find out which factors or attributes are important for the students' retention. It is important because higher retention rate will increase university funding and also help them to recruit more students and faculties in order to give a better service to the community. College student graduation rates are often used as a measure of institution's performance. These kinds of analyses play a significant role for universities' decision about how to expand their majors' capacities and how to allocate financial aid between students.

2.1 Literature Review

Event prediction is an important area of research where the goal is to predict the occurrence of an event in the data [18]. In higher education, many modeling techniques were found to help educational institutions to predict at-risk students [28]. This results in planning for interventions and better understanding and addressing fundamental issues that cause the student attrition problem. 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 find the important factors that have some effect on this prediction [22, 43]. 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 [8, 24]. [25] 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 [15].

Recently, many researchers in the area of machine learning and data mining tried to address the student retention phenomenon in college and university [10, 35, 41]. 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 [3, 29], Decision tree [17, 32, 42], Boosting methods and support vector machines [44] have been developed to predict student attrition rate with higher accuracy compared to the traditional statistical methods.

A slightly more complex and relevant modeling technique is survival analysis. 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 [23, 26]. In this type of regression model, both components, (i) if an event (i.e. dropout) occurs or not and (ii) when the event occurs can be incorporated simultaneously [27]. So we can assign the probability of dropout for a single period of time and we can also assign a probability for each time period (e.g., semesters) [9]. Thus, the benefit of using survival analysis over logistic regression or other data mining methods is the ability to add time component to the model and also effectively handling censored data. However, the literature in this area is limited. The use of survival analysis to study both student retention and student dropout has been developed in [19, 20, 21]. Among those, only [19] developed an event history mode to assess the attrition behaviour among first-generation students using pre-enrollment attributes using linear hazard rate. However, they did not use static and time-dependent variables simultaneously in a more complex survival model such Cox proportional hazard model to find non-linear relation between the attributes in student retention problem.

Hence, reviewing the literature, despite the fact that survival model has more flexibility to handle the student retention problem, there were only a few efforts in this domain to model the student dropout data. Therefore, it is evident that there is considerable room for improvement in the current state-of-the-art. This thesis will further the existing work related to the student success by showing an in-depth application of both static and time-dependent 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 before in the literature.

2.2 Prediction Models

In such longitudinal data, event prediction is an important area of research where the goal is to predict the occurrence of an event [18]. In this section, different regression and classification methods will be introduced which are widely used in other domains. These methods have also been used in student dropout prediction. The performance of these methods will be compared to our proposed methods in chapter 4.

2.2.1 Regression

Regression models are among the most important methods in predictive analytics [12]. They have been widely used in many domains of studies. Regression methods are primary established to model a mathematical equation to represent the interactions between the different variables in consideration. Depending on the situation, there are a wide variety of models that can be applied. The most popular one is linear regression which tries to find the linear relation between dependent variable and a set of independent or predictor variables.

$$y = w_0 + \sum_{k=1}^{p} w_k x_k \tag{2.1}$$

where w_k is the coefficient of the k^{th} variable. The goal here is to select the parameters of the model so as to minimize the sum of the squared error. In the student retention problem, it was one of the primary techniques that has been used before [11]. It can be applied to model time-to-event problem, which in the student data is the time to dropout. However, it has some drawbacks that will be discussed in more detail in the next chapter.

2.2.2 Logistic Regression

One of the well-established statistical model is the Logistic Regression where the dependent variable is categorical [40]. In this model, logit transformation of a linear combination of the attributes is used to resolve a binary classification problem. For example, consider X and its k^{th} feature as x_k, then Y is the predicted output and in our case it is binary. If each example has a label to be either -1 or +1, and there are p number of features in each instance, the model has the following form

$$\log \frac{Pr(y = +1 \mid x)}{Pr(y = -1 \mid x)} = \sum_{k=0}^{p} w_k x_k = z \tag{2.2}$$

here, $x_0 = 1$ is an additional feature called “bias”, and w_0 is the corresponding “bias weight”. From Eq. (2.2), we have

$$Pr(y = +1 \mid x) = \frac{e^z}{1 + e^z} = g(z) \tag{2.3}$$

Logistic regression models are usually fit by maximizing the log-likelihood function

$$L(w) = \sum_{i=1}^{n} logPr(y = y_i \mid x_i) = \sum_{i=1}^{n} logPr(y_i z_i) \tag{2.4}$$

where n is the number of instances and $z_i = \sum_{k=0}^{p} w_k x_{ik}$. To solve this maximization problem, it is common to use Newton’ s method.

2.2.3 Support Vector Machines (SVM)

Support vector machines (SVMs) are supervised learning models used for classification and regression analysis [4]. SVM constructs a hyperplane or a set of hyperplanes in a high-dimensional space, which can be used for classification (SVC) or regression (SVR).

In the linear SVC, the goal is to optimize

$$\text{minimize } \frac{1}{2}|w|^2$$

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1$$

where the y_i is either 1 or 1, indicating the class to which the point $\mathbf{x}_i$ belongs. Each $\mathbf{x}_i$ is a p-dimensional real-valued vector. The goal is to find the maximum-margin hyperplane that divides the points having y_i=1 from those having y_i=-1.

Support Vector Regression (SVR) is a sub-category of SVM where it can solve regression problems [13]. The model produced by Support Vector Regression depends only on a subset of the training data. Thus, training the original SVR means solving

$$\text{minimize } \frac{1}{2}|w|^2$$

having these two constraints, $\begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \end{cases}$

where in the above optimization problem x_i is a training sample with target value y_i and ϵ is a threshold parameter.

2.2.4 Adaptive Boosting

Adaptive Boosting algorithm, also known as Adaboost, is a widely used machine learning technique that was first introduced in 1997 [14]. The idea behind it is to create highly accurate predictive model by combining other learning algorithms, called “weak learners” to improve their performance. By weighting the sum of output of those learned, the final result is a boosted classifier. A boosted classifier is a classifier in the form

$$F_T(x) = \sum_{t=1}^{T} w_t f_t(x) \tag{2.5}$$

where w_t is the weight of classifier and f_t is a weak learner that takes an object x as input and returns the class of the object. In the training phase, based on the sign of the weak learner, we can identify the class of predicted object and the absolute value gives the confidence in that classification. Each weak learner produces an output, hypothesis $h(x_i)$ for each sample in the training set. At each iteration t, a weak learner is selected and assigned a coefficient α_t such that the sum of training error E_t of the resulting t-stage boost classifier is minimized. This algorithm has a weighting phase, where at each iteration of the training process, a weight is assigned to each sample in the training set equal to the current error on that sample. These weights are used to learn the training of the weak learner.

2.2.5 Decision Tree

A decision tree is one of the powerful predictive machine learning model that decides the target value of a new sample based on various features of the data. The internal nodes of a decision tree denote the different attributes, the branches between the nodes indicate the possible values that these attributes could have in the observed samples, while the terminal nodes will give the final value of the dependent variable. One of the important algorithms in this field is C4.5, which is used to generate a decision tree, was developed by Ross Quinlan [33]. C4.5 builds decision trees from a set of training data using the concept of information entropy. Entropy is a measurement of uncertainty in any random variable. In C4.5, at each node of the tree, the algorithm chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is based on entropy. Entropy for any node t can be written as

$$Entropy(t) = -\sum_{j} p(j|t) log p(j|t) \tag{2.6}$$

where $p(j|t)$ is the relative frequency of class j at node t. So, the idea of C4.5 is to measure the reduction in Entropy achieved because of the split. The information gain for any parent node p can be measured as

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} Entropy(i) \right) \tag{2.7}$$

where parent node p split in to k partitions and n_i is number of records in partition i. Consequently, the attribute with the highest information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

We apply the methods that have been introduced in this chapter on the student retention data for comparison to our proposed survival analysis methods which will be introduced in the next chapter.

CHAPTER 3

PROPOSED SURVIVAL ANALYSIS APPROACH

Survival models are used to estimate time to events of interest. The ability to model the dynamic nature of incidents is a powerful tool because in many cases answer to the question of *"when"* is as important as *"who"*. Furthermore, it is important to identify the characteristics that led to the occurrence of event. This is very important in student retention problem as retention is not an instant event, but rather a lengthy process that totally depends on time [39]. Hence, survival analysis methods would be an appropriate choice to model this kind of problem.

The main objective of this chapter is to explain existing survival methods and utilize them for the student retention problem. Survival analysis methods have been developed in the field of statistics to handle data of time to some event or failure. However, little research in higher education has focused on using the survival methods for predicting retention. We also explore the use of time-dependent covariates and the exciting opportunities that it offers. Basically, in this chapter, we try to answer the three important questions in the student retention problem:

- Which factors are significant for student dropout?
- Does the student dropout or not? How will the time-depended factors affect that?
- What is the risk of dropout at each semester? In other words, in which semester the dropout will happen?

To answer all these questions, we aim to build survival analysis method which can be used to identify student at risk and predict the probability of dropout at each semester while considering time-dependent attributes like GPA for each semester.

3.1 Survival Analysis

Survival analysis is as a set of techniques that can be used to analyze data where the outcome variable is the time until the occurrence of an event of interest. This kind of data has three main characteristics: (1) the dependent variable (or response) is the time until the occurrence of an event, (2) the time for observations are censore i.e., for some subjects the event of interest has not occurred (or not recorded) at the time the data is analyzed, and (3) there are predictors or attribute variables that have effect on the time to event.

One of the important characteristics of longitudinal data is that, it can be incomplete due to the inability to continuously track the subject, also referred to as *censoring.* This incompleteness in events or information in longitudinal data is in many ways different from missing data problems encountered in routine data mining problems, and not all modeling techniques are able to handle them. Thus, it becomes difficult for standard machine learning methods to model data which contains censoring. Ignoring the censored data on one hand yields suboptimal biased models 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. For the student retention problem, a question that arises is that why linear regression cannot be used to model semester of dropout. One important reason is that linear regression cannot handle the censored observations in an efficient way. Unlike ordinary regression models, survival models incorporate information from both uncensored and censored data to evaluate important features. One of the paramount aspects of survival analysis is covariate data, which are collected longitudinally. It appears to be regular and proper to use the covariate information that varies over time in an appropriate statistical model. This aspect of time-dependent covariates make survival data very unique in a way that other standard machine learning methods could not handle.

In spite of the success of survival analysis methods in other domains such as healthcare, there is only a limited attempt of using these methods for the student retention problem [31, 34]. In this section, after some basic definitions, the proposed survival analysis framework for student dropout prediction will be explained. Before that the notations used in this thesis will be introduced in Table 3.1.

Table 3.1: Notations used in this thesis.

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 feature vectors for subject i
$Z_i(t)$	$1{\times}q$ matrix of time dependent feature vectors for subject i
T	$n \times 1$ vector of event times
C	$n \times 1$ vector of last follow up time
O	$n \times 1$ vector of observed time which is $min(T, C)$
δ	$n \times 1$ binary vector of censored status
d_i	number of events occurred at time t_i
$S_0(t)$	base survival probability
$S(t \mid X, Z(t))$	conditional survival probability at time t
$h_0(t)$	base hazard rate
$h(t \mid X, Z(t))$	conditional hazard probability
β	$p \times 1$ vector of Cox regression coefficient
$L(\beta)$	maximum likelihood function for β

3.1.1 Survival and Hazard Functions

Survival analysis consist of two main components: first is the event time and the second one is the status of the event which has the occurrence information for the event of interest. With event time, we can fit it into two functions that are dependent on time, namely the survival and the hazard functions. These two functions are critical concepts in survival analysis to describe the distribution for times of events. For every specific time the survival function gives the survival probability until that time. The hazard function gives the possibility that the event will occur, per time unit.

Let T denotes the survival time of an individual, which has density f. The density f and the distribution function $F(x) = \int_0^x f(u)\,du$ are not particularly informative about the chance of survival at a given time point. Instead, the survival, hazard, and cumulative hazard functions, which are functions of the density and distribution functions, are used.

Survival Function:

It is defined as the probability that the event of interest has not occurred by t. The survival function can be expressed in terms of probability distribution and probability density functions:

$$S(t) = Pr\{T > t\} = \int_t^\infty f(u)\,du = 1 - F(t) \tag{3.1}$$

Hazard Function:

An alternative characterization of the distribution of T is given by the hazard function, or instantaneous rate of occurrence of the event, defined as

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T < t + dt)}{dt} \tag{3.2}$$

In other words, $h(t)$ is defined as the event rate at time t conditional on survival until time t. The numerator of this expression is the conditional probability that the event will occur in the interval $[t; t + dt)$ given that it has not occurred until time 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 to zero, we obtain an instantaneous rate of occurrence.

The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval $[t; t + dt)$ and $T \geq t$, to the probability of the condition $T \geq t$. The former may be written as $f(t)dt$ for small dt, while the latter is $S(t)$ by definition. Dividing by dt and passing to the limit gives the useful result

$$h(t) = \frac{f(t)}{S(t)} \tag{3.3}$$

In other words, the rate of occurrence of the event at duration t equals the density of events at t, divided by the probability of surviving to that duration without experiencing the event. $S(t)$ can also be expressed as

$$S(t) = exp\left(-\int_0^t h(x)\,dx\right) \tag{3.4}$$

Cumulative Hazard Function:

It is defined as the sum of the risks that someone faces going from duration 0 to t. These results show that the survival and hazard functions provide alternative but equivalent characterizations of the distribution of T.

$$H(t) = \int_0^t h(x)\,dx \tag{3.5}$$

3.1.2 Censored Data

One of the main features of survival data which distinguishes it from all other kinds of data, is that it is often incomplete. This means that the event information for some observations is not complete and such instances are considered to be censored. There are three main different types of censoring; right, left and interval censoring. Most of censoring are right which means we may not observe the time of event occurrence, and only have knowledge that the individual survived until a certain time point. This means for some reason independent of its survival time, the individual chooses to leave the study. In student retention problem if the event of interest is "dropout" then the type of censoring we will consider will be right censoring.

Let us suppose that T_i is the survival time, but this may not be observed and we observe instead $Y_i = min(T_i, C_i)$, where C_i is the censoring time. We do know that, if the data has been censored, and together with Y_i we observe the indicator variable

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

So, if for individual i, $\delta_i = 0$, it is censored and if $\delta_i = 1$ it is not censored. Figure 3.1 illustrates the student retention problem using survival analysis in which students A, B and D dropout before semester 6 and students C, E and F remain at school by the end of the 6^{th} semester or in other words they are censored at semester 6 (shown with X in the figure).

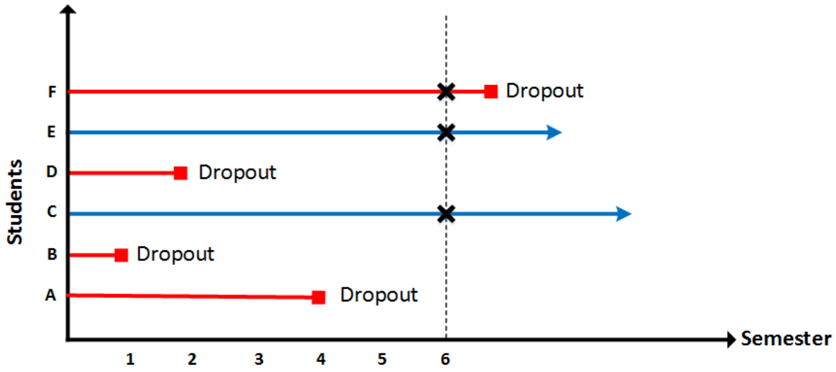


Figure 3.1: Illustration of the student retention data using survival concepts.

3.1.3 Maximum Likelihood Function

In order to estimate parameters or making other kinds of inferences for survival models, they can be viewed as ordinary regression models in which the response variable is time. However, the likelihood function in the presence of censored data is more complicated. The likelihood function for a survival model will be a mixture of probabilities and densities, depending on whether the observation was censored or not. By definition, the likelihood function is the product of the likelihood of each individual. It is convenient to partition the data into four categories: uncensored, left censored, right censored, and interval censored. These are denoted by “unc.”, “l.c.”, “r.c.”, and “i.c.” respectively in the equation below and, in general, it can be formulated as follows:

$$L(\theta) = \prod_{T_i \in unc.} \Pr(T = T_i \mid \theta) \prod_{i \in l.c.} \Pr(T < T_i \mid \theta) \prod_{i \in r.c.} \Pr(T > T_i \mid \theta) \prod_{i \in i.c.} \Pr(T_{i,l} < T < T_{i,r} \mid \theta) \tag{3.6}$$

In this thesis we only consider data that event occurred for them and right censored data. Then above likelihood function can be written as:

For uncensored data, we have

$$\Pr(T = T_i \mid \theta) = f(T_i \mid \theta)$$

For right-censored data, we have

$$\Pr(T > T_i \mid \theta) = 1 - F(T_i \mid \theta) = S(T_i \mid \theta)$$

3.1.4 Non-parametric and Parametric Survival Models

The analysis of survival data can be done in multiple ways. One of the common methods is non-parametric where there is no assumption about the form of the survival distribution. One of the well-known non-parametric estimator of the survival function is the Kaplan Meier method which is widely used to estimate and graph survival probabilities as a function of time. It can be used to obtain univariate descriptive statistics for survival data, including the median survival time, and compare the survival experience for two or more groups of subjects. This estimator is defined as

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.7}$$

where d_i represents the number of failures at time t, and n_i indicates the number of individuals who have not experienced the event of interest, and have also not been censored, by time t.

On the other hand, parametric methods assume that the underlying distribution of the survival times follows a certain known probability distribution. Popular ones in this category include the exponential, Weibull, and Lognormal distributions. The description of the distribution of the survival times and the change in their distribution as a function of predictors is of interest. Model parameters in these settings are usually estimated using an appropriate modification of the maximum likelihood function.

3.2 Cox Proportional Hazard Model

In previous section, we discussed two categories of techniques for survival data modeling. Non-parametric method only considers time-to-event data and does not take care of any covariate information that might relate to the event occurrence. On the other hand, parametric models such as Accelerated Failure Time (AFT) are able to consider covariates in the model however, we should know the distribution that the data follows. There is another category of methods, referred to as semi-parametric, which we can model survival data using covariate while we do not need to make any specific assumption for the time-to-event data. One of the popular method in this category is Cox proportional hazard model [7] which makes fewer assumptions than typical parametric methods but more assumptions than non-parametric methods [5]. In particular, and in contrast with parametric models, it makes no assumptions about the shape of the baseline hazard function [6].

Let T_i denote the observed time that can be either censoring time or event time for subject i, and let δ_i be the event status indicator; if $\delta_i = 1$, then the event occurred and if $\delta_i = 0$, then the subject is censored. The hazard function for the Cox proportional hazard model has the form

$$h(t|X) = h_0(t)\exp(\beta_1 X_1 + \cdots + \beta_p X_p) = h_0(t)e^{(\beta X)} \tag{3.8}$$

where $h_0(t) = e^{\alpha(t)}$ is the baseline hazard function at time t and $\exp(\beta_1 X_1 + \cdots + \beta_p X_p)$ is the risk associated with the covariate values. If we take the ratio of the hazards, the baseline hazard cancels out and the hazards are proportional at any given time t, yielding the proportional hazards model. This expression gives the hazard at time t for an individual with covariate vector X. Therefore, the survival probability function for Coxph model can be formulated as

$$S(t \mid X) = S_0(t)^{exp(\beta X)} \tag{3.9}$$

where

$$S_0(t) = e^{-\int_0^t h_0(x)dx} \tag{3.10}$$

3.2.1 Parameter Estimation

Parameter estimation in the Coxph regression model is done by maximizing the partial likelihood as opposed to the likelihood. If n defines the number of subjects, $f_i(t_i)$ the density function of the failure, $S_i(t_i) = P(T_i > t)$ is the survival function, t_i is the minimum of the exact failure time T_i and the censoring time C_i of the i^{th} individual and $\delta_i = I(T_i \leq C_i)$ is an indicator variable which represents the failure status, then the maximum likelihood function contains two parts as below:

$$L(\beta) = \prod_{i=1}^{n} [f_i(t_i)]^{\delta_i} \times [S_i(t_i)]^{(1-\delta_i)} \tag{3.11}$$

If the event has occurred for the individual then, $\delta_i = 1$, and the second term will be zero and we only have $f_i(t_i)$ for it. On the other hand, if the event does not happen then we only care about the second part which will give the probability that an individual survives over time. Also, we know that $h_i(t_i) = \frac{f_i(t_i)}{S_i(t_i)}$ and defined as hazard function at time t_i, then Eq. (3.11) changes to

$$L(\beta) = \prod_{i=1}^{n} [h_i(t_i)S_i(t_i)]^{\delta_i} \times [S_i(t_i)]^{(1-\delta_i)} \tag{3.12}$$

and finally we will have

$$L(\beta) = \prod_{i=1}^{n} [h_i(t_i)]^{\delta_i} \times S_i(t_i) \tag{3.13}$$

Log likelihood function, $l(\beta)$, can be found

$$l(\beta) = \sum_{i=1}^{n} \delta_i \times [h_i(t_i)] + \sum_{i=1}^{n} S_i(t_i) \tag{3.14}$$

Based on Cox regression formula, a partial likelihood can be constructed from the data as follows:

$$L(\beta) = \prod_{i:\delta_i=1} \frac{\theta_i}{\sum_{j:t_j \geq t_i} \theta_j} \tag{3.15}$$

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

$$\hat{h}_0\big(t_{(i)}\big) = \frac{1}{\sum_{j \in t_{(i)}} \theta_j} \tag{3.16}$$

where $t_{(i)}$ is the time order for event occurrence such that $t_{(1)} < t_{(2)} < ... < t_{(n)}$. The proportional hazard condition states that covariates are multiplicatively related to the hazard. In the simplest case, the precise effect of the covariates on the life-time depends on the type of $h_0(t)$. The Cox partial-likelihood, shown above, is obtained by using the Breslow's estimate of the baseline hazard function, plugging it into the full likelihood and then observing that the result is a product of two factors. The first factor is the partial likelihood, in which the baseline hazard has cancels out. The second factor is independent of the regression coefficients and depends on the data only through the censoring pattern. The effect of covariates estimated by any proportional hazards model can thus be reported using hazard ratio.

3.3 Time-Dependent Cox Model (TD-Cox)

The Cox proportional hazard model makes an assumption that covariates are independent of time. In other 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 the student retention problem) where the covariates change over time and the above assumption does not hold. Thus, it is more appropriate to use time-dependent covariates which will potentially result in more accurate estimates of the outcomes.

Consequently, we can define time-dependent variables that can change in value over the course of the observation period. Variables such as body weight, income, marital status or student GPA are few examples of attributes that could vary over time. One way to look at the time varying covariates is to hold the values of such variables fixed at a certain point in time, say baseline, but to have an accurate analysis the best way is to change variables over the time. To extend the logged hazard function to include variables that change over time, for each time varying covariate in the model, we can represent it as a function of t. Thus, Cox proportional hazard model can be written as

$$h(t|Z(t)) = h_0(t)\exp(\beta_1 Z_1(t) + \cdots + \beta_q Z_q(t)) = h_0(t)e^{(Z(t)\beta')} \tag{3.17}$$

So, this function now means that the hazard at time t depends on the value of Z at time t. Extensions to time varying attributes can be incorporated using the counting process formulation [2]. Essentially, in the counting process, data are expanded from one record-per-subject to one record-per-interval between each event time for each subject. Covariate information needs to be updated and available at these times, but not in between. Algorithm 1 outlines the reformatting process for time dependent survival data using counting process.

Algorithm 1 *Reformatting Time Dependent Survival Data Based on Counting Process*

```
Require: Survival data $D_n = (X, Z(t), T, \delta)$
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 $j = Tc$ then
                $s_{i+j}$=1
            else
                $s_{i+j}$ =0
            end if
        end for
    end for
end for
return reformatted data $D = (X, Z, t, s)$
```

In other to have a better understanding of counting process, we demonstrate it using an example. Table 3.2 shows the data record-per-student format. Using Algorithm 1, data changes to record-per-interval between each event time (Table 3.3), per student. In this example, for each student, we record time of dropout and status. If status is 1, it means student dropout and if 0 it means student does not dropout until the observed time. Also, we obtain GPA for each semester. In order to do time-varying survival analysis, we should change the format using counting process. Table 3.3 shows the result of reformatting. Basically, we consider the time interval by adding t_0 column and for each interval, GPA is shown separately. Other static variables such as demographic information can be also added without changing over intervals.

Table 3.2: Example of survival data

Student ID	**time**	**status**	**GPA(t=1)**	**GPA(t=2)**	**GPA(t=3)**
ID_1	1	1	2	-	-
ID_2	2	1	3.2	1.8	-
ID_3	3	0	4	4	3.5

Table 3.3: Example of survival data after counting process 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 thesis, we use the Time Dependent Cox proportional hazard function, namely TD-Cox, which contains a mixture of static and time varying covariates. Thus, the hazard function can be defined as

$$h(t|X, Z(t)) = h_0(t)\exp(\beta_1 X_1 + \cdots + \beta_p X_p + \beta_{(1+p)} Z_1(t) + \cdots + \beta_{(p+q)} Z_q(t)) = h_0(t) e^{(X+Z(t))\beta')} \tag{3.18}$$

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

$$S(t \mid X, Z(t)) = S_0(t)^{exp(\beta(X+Z(t)))} \tag{3.19}$$

where $S_0(t)$ can be estimated using Eq. (3.10). Algorithm 2 summarizes the *TD-Cox* method. First we need to transform data using Algorithm 1. In line 2, we learn TD-Cox parameters using training data using maximum likelihood function explained in section 3.4. Then for each test data we use Eq. (3.19) to estimate survival probability and event probability (lines 2-5). Figure 3.2 summarized the survival analysis framework that we developed for student retention problem.

Algorithm 2 *TD-Cox method*

```
Require: reformatted data $D = (X, Z, t, s)$ from Algorithm 1
1: Learn TD-Cox parameters, $\beta$ and $\hat{h}_0$ Eq. (3.15) and Eq. (3.16)
2: for all students in test data do
3:    Estimate $\hat{S}(t \mid X, Z)$ from Eq. (3.19)
4:    $\hat{F}(t \mid X, Z) = 1 - \hat{S}(t \mid X, Z)$
5: end for
6: return probability of event occurrence
```

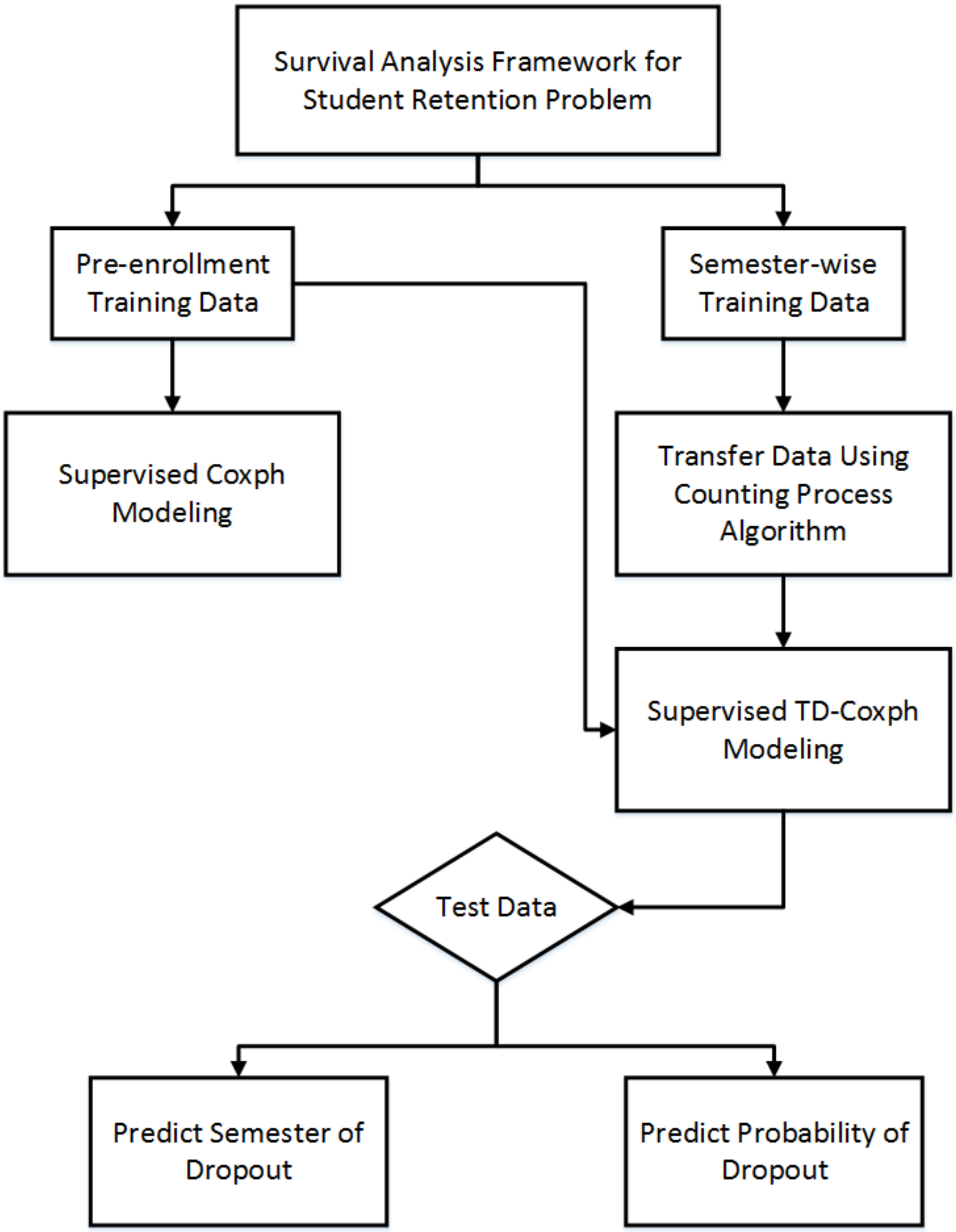

Survival Analysis Framework for Student Retention Problem
Pre-enrollment Training Data
Semester-wise Training Data
Supervised Coxph Modeling
Transfer Data Using Counting Process Algorithm
Supervised TD-Coxph Modeling
Test Data
Predict Semester of Dropout
Predict Probability of Dropout

Figure 3.2: Flowchart of proposed survival analysis framework for student retention problem

CHAPTER 4

EXPERIMENTAL RESULTS

In this chapter, we present the results of the proposed survival analysis framework for predicting student dropout at Wayne State University. First, we explain our data source and define the variables used in our model along with their descriptive statistic. We also discuss the evaluation method to check the performance of the proposed method. We show the experimental result for two types of analysis: “predicting student dropout” and “estimating the semester of dropout”. Finally, the practical implications of our framework in educational studies will be also discussed.

4.1 Data Description

The data for the student dropout prediction analysis has been collected from the Wayne State University (WSU) database. They are distributed across various sources. Each of these sources provides distinct information about each student. We extracted and summarized the information we collected from these different sources. In the student database, all the data for students who are admitted to Wayne State since 2002 is available. We only focus on FTIAC (First Time in Any College) students and the transfer students are not considered in our dataset. The main reason for not considering these transfer students is that the duration of study for a transfer student is different from FTIAC students because their graduation pattern is quite different than other students. Thus, we use FTIAC students in all majors and colleges from 2002 to 2009 which come to a total of 11,121 students. We consider data from 5 colleges: fine art, liberal art and science, education, engineering and business school. We did not consider school of medicine and college of nursing because their patterns are different. In order to have a better understanding of coefficients that are estimated by our model, all the data were standardized. After all the preparation and necessary pre-processing, we ended up with

33 attributes which could be categorized into six different groups:

- Demographic,
- Family Background,
- Financial Attributes,
- High School Attributes,
- College Enrollment Attributes,
- Semester-wise Attributes.

Among these categories, only Semester-wise group attributes are time-dependent and we use them further for time-dependent survival analysis. The details of all the attributes used in our analysis is summarized in Table 4.1. We will now define some of the terms that will be used in this chapter.

- ***Dropout Student***: It is defined as a student who does not register in a semester or whose semester GPA is zero.
- ***Event***: The student dropout before graduation is our event of interest.
- ***Censored***: If a student does not dropout within first 6 semesters or we have no information about it, then it is defined as censored data.
- ***Status***: It is 1 if student dropout within first 6 semesters and 0 if he continues study and never drops out in that period.
- ***Time***: Semester in which the dropout occurred for a given student.

In order to evaluate the performance of 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 up to first 6 semesters. The illustration of this experiment shows in Figure 3.1.

- *Experiment 2:* In this experiment, we are not following all the student for 6 semesters. In other words, we cut the observation at 2009, so in this case, for students who have been admitted to school in 2008, we have records for only two semesters. For a better understanding, we illustrate this experiment in Figure 4.1. As it is shown, in this experiment we have censored data before 6^{th} semester.

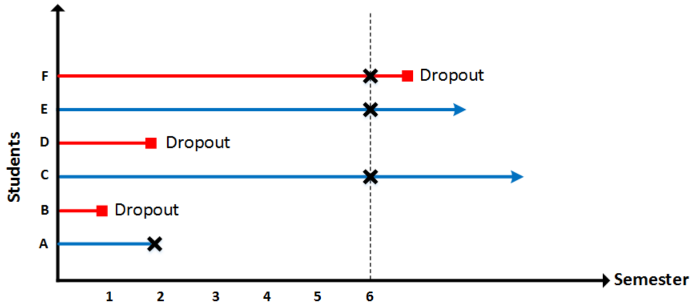


Figure 4.1: An illustration to demonstrate second experiment.

4.2 Preliminary Analysis of the Data

In order to have a better understanding of the WSU student retention data, in this section, we provide some descriptive analysis. Let us start with Figure 4.2 which shows the percentage of dropout in each semester. It indicates the importance of addressing dropout issue as early as possible in the student school life. From the figure, the percentage of first year dropout is around 35% and by end of second year it increase up to 55%.

Figure 4.3 shows histogram plot for some of the important pre-enrollment attributes. It is clear that high school GPA and ACT score significantly affect the student dropout within the first 6 semesters. The higher the GPA and ACT score, the less chance of

Table 4.1: Description of attributes used to build our dataset.

Attribute		Type	Description
Demographic Attributes			
1	Gender	Binary	Male or Female
2	Marital status	Binary	Married or Single
3	Ethnicity	Categorical	Ethnicity include White, Black, Asian, etc.
4	Hispanic or non-Hispanic	Binary	Hispanic or non-Hispanic
5	County code	Categorical	County code includes Wayne, Oakland, Macomb and other
Family Background Attributes			
6	Father's education level	Categorical	Education level of father
7	Mother's education level	Categorical	Education level of mother
8	Number of family members	Numeric	Number of Family members
9	Number of family members in college	Numeric	Number of family members graduated from the college
Financial Attributes			
10	Student cash amount	Numeric	Amount of available cash from student
11	Student's parents cash amount	Numeric	Amount of available cash from parents
12	Student income	Numeric	Student's income from work
13	Father's income	Numeric	Father 's income from work of the student
14	Mother's income	Numeric	Mother 's income from work of the student
15	Parent 's Income	Numeric	Parent's income
High school Attributes			
16	High School GPA	Numeric	GPA in high school
17	Composite ACT score	Numeric	Score of Composite ACT
18	Math ACT score	Numeric	Score of Math ACT
19	English ACT score	Numeric	Score of English ACT
20	Reading ACT score	Numeric	Score of Reading ACT
21	Science ACT score	Numeric	Score of Science ACT
22	High school graduation age	Numeric	Student's graduate age from high school
College Enrollment Attributes			
23	Age of admission	Numeric	Age when the student was admitted to the University
24	First admission semester	Categorical	First semester of admission
25	Degree awarded or not	Binary	Whether student gets a degree or not
26	Credit transferred or not	Binary	Number of transferred credits
27	Major	Categorical	Department information
28	College	Categorical	One of the five colleges considered
Semester-wise Attributes			
29	Credit hours attempts	Numeric	Number of credits taken by student
30	Percentage of passed credits	Numeric	Fraction of credits passed by student
31	Percentage of dropped credits	Numeric	Fraction of credits dropped by student
32	Percentage of failed credits	Numeric	Fraction of credits failed by student
33	GPA	Numeric	GPA of specific semester

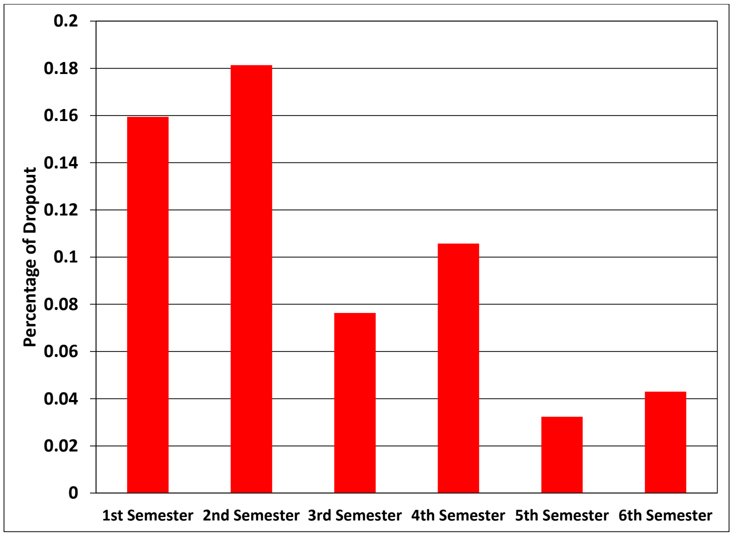
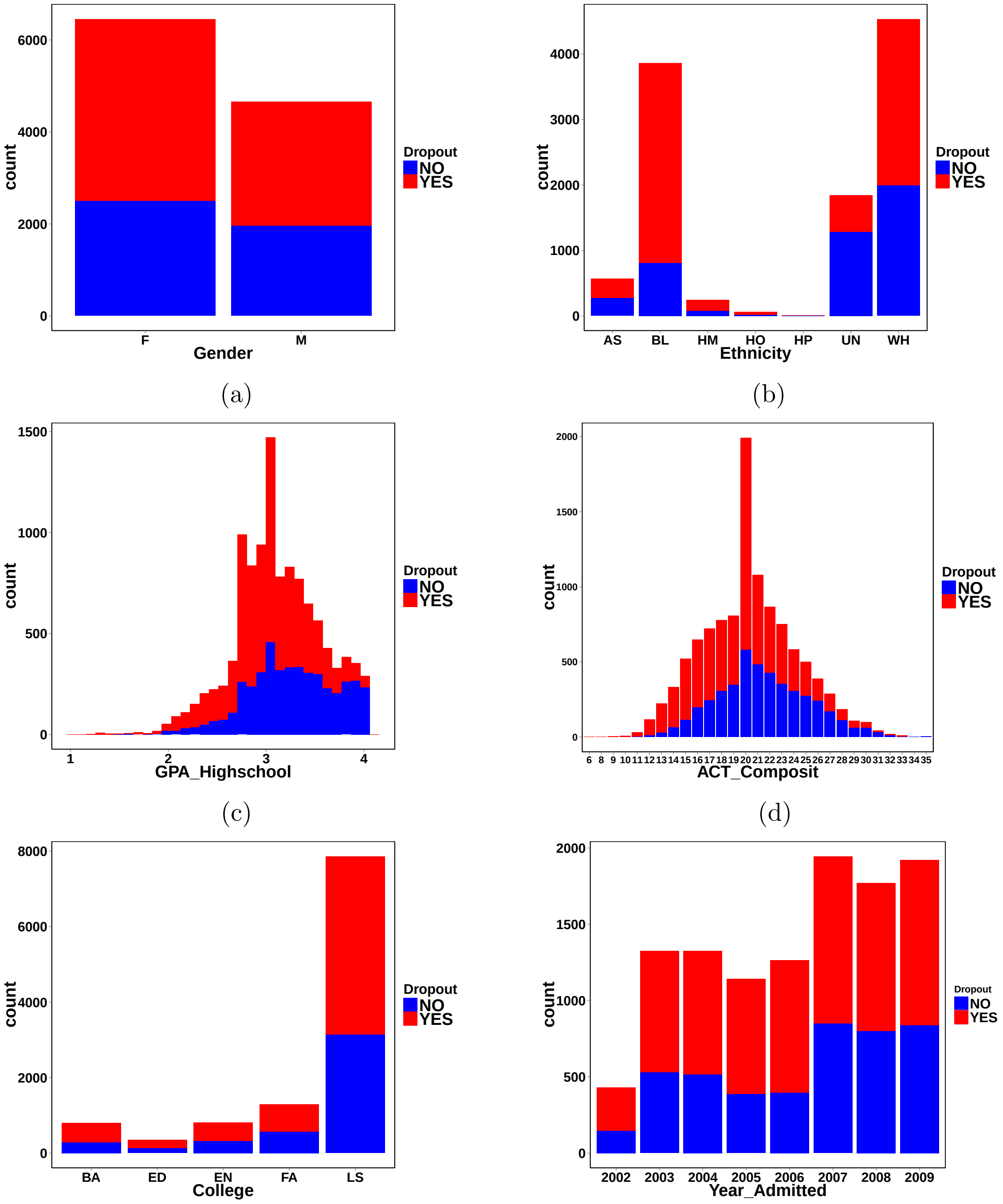


Figure 4.2: Histogram plot for the percentage of dropout per semester

dropout from school. Also the gender and ethnicity could be good indicators for student at risk of dropout.

We also did an exploratory survival analysis using Coxph estimator in order to see the dropout pattern for different scenarios. Figure 4.4 shows the probability of dropout for 4 different scenarios. In each scenario, we consider all other variables which are set at average level and allow only one attribute to have a different value. As an example, in Figure 4.4(a), we test the probability of dropout for male vs. female. It is clear that gender does not have significant impact on student dropout by itself. Figure 4.4(b) shows the effect of different county residency on student dropout. It indicates that those who come from the three nearby counties have a lower chance of dropout. In Figure 4.4(c) we test the impact of different ethnicity on student dropout. Finally Figure 4.4(d) shows the importance of high school GPA on student dropout.



(a)

(b)

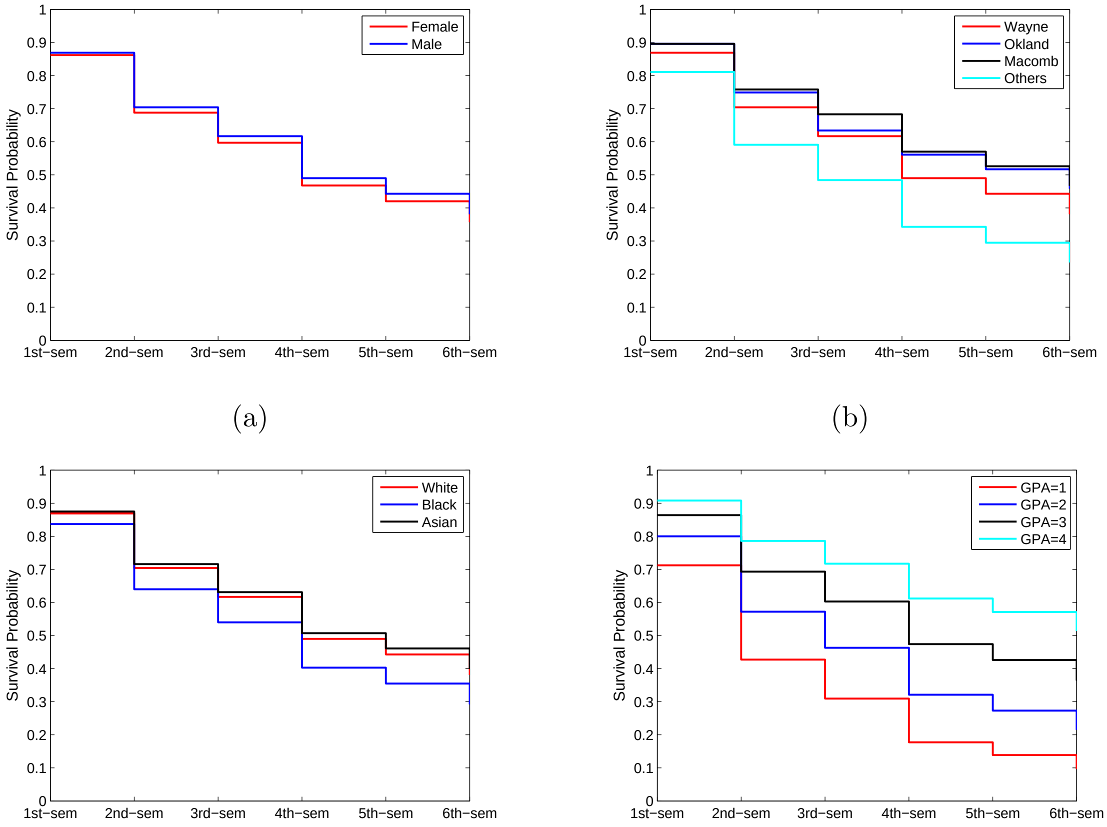
(c)

(d)

(e)

(f)

Figure 4.3: Histogram plot for some of the important pre-school attributes: (a) Gender, (b) Ethnicity, (c) High school GPA, (d) ACT composite score, (e) College and (f) Year of admission



(a)

(b)

(c)

(d)

Figure 4.4: Probability of attrition for 4 different variables: (a) gender, (b) county of residency, (c) ethnicity and (d) high school GPA. It should be noted that all other variables fix on an average for each case.

4.3 Evaluation Metrics

In order to have a quantitative measure of 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. Then our test data consists of student admitted in 2009 and they are completely unused during our 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 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 [36] and for the rest of the methods we used open source Weka software [16]. To assess the performance of the proposed model, the following metrics are used for the classification problem.

- ***Accuracy*** is expressed in percentage of subjects in the test set that were 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 is the true positive, FP is false positive and FN is false negative.

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

For the time to dropout prediction which is originally a regression problem, we used following metrics:

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

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

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

- ***Root Mean Square Error (RMSE)*** is a frequently used measure of the differences between values predicted by a model and the values actually observed. The RMSE represents the sample standard deviation of the differences between predicted values and observed values.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{n}}$$

- ***Relative Absolute Error (RAE)*** measures the size of the error in percentage terms as follows

$$\text{RAE} = \frac{\sum_{i=1}^{n} |\hat{y_i} - y_i|}{\sum_{i=1}^{n} |\bar{y} - y_i|}$$

where $\bar{y}$ is the average of actual values.

- ***Root Relative Square Error (RRSE)*** is similar to the RAE for root mean square error which can be calculated as

$$\text{RRSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2}}$$

4.4 Performance of Various Classification Methods

In this section, we demonstrate the performance of the proposed survival analysis method, *Coxph* and *TD-Cox*, on Wayne State University student dropout information from 2002 to 2009. Table 4.2 shows the list of attributes along with their coefficients from the Coxph model. In this table, **coef** shows coefficients, **exp(coef)** shows the exponential of each coefficient and **se(coef)** is standard error of the coefficients. The **z** value is the Wald statistic for testing the hypothesis that the coefficient is zero and **Pr(**$>|z|$**)** is the tail area in a 2-tail test. The results show that the number of family members, number of family member at college, county of residency, age of admission to college, ethnicity, high school GPA and ACT score have a significant impact on student dropout.

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, namely, 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 4.3-4.6. From these Tables, we can see the consistent results of the *TD-Cox* method which beat all other methods. Comparing results in Tables 4.3 and 4.4 with Tables 4.5 and 4.6, it is clear that we could get higher performance from *Cox* and *TD-Cox* (survival analysis methods) in the presence of more censored data. In this study, as described in Table 4.1, we define 5 after-enrollment variables including GPA, percentage of passed,

Table 4.2: Coefficient estimation of the attributes from the Coxph model

Attributes	**coef**	**exp(coef)**	**se(coef)**	**z**	**Pr**($>\|z\|$)
Father_Edu	-0.017175	0.982972	0.015994	-1.074	0.28289
Mother_Edu	-0.027029	0.973333	0.017743	-1.523	0.12767
No_Family_Member	0.094209	1.098789	0.010161	9.272	2.00E-16
No_At_College	-0.566458	0.567532	0.024475	-23.145	2.00E-16
County_Okland	0.002401	1.002404	0.044641	0.054	0.95711
County_Other	0.644733	1.905478	0.061575	10.471	2.00E-16
County_Wayne	0.23978	1.270969	0.03684	6.509	7.58E-11
Transfer_Credit	-1.009215	0.364505	0.042285	-23.867	2.00E-16
CollegeED	-0.041782	0.959079	0.082031	-0.509	0.61051
CollegeEN	-0.098913	0.905821	0.064003	-1.545	0.12224
CollegeFA	-0.166346	0.846754	0.058515	-2.843	0.00447
CollegeLS	-0.046164	0.954886	0.046968	-0.983	0.32567
GPA_Highschool	-0.419755	0.657208	0.031141	-13.479	2.00E-16
Age_Highschool_grad	-0.006037	0.993981	0.011206	-0.539	0.59008
Age_Enter_College	0.013679	1.013772	0.003677	3.72	0.0002
Gender	-0.063127	0.938824	0.027993	-2.255	0.02413
Marrital_StatusS	0.004511	1.004521	0.095535	0.047	0.96234
Hispanic_YES	-0.829685	0.436187	0.31907	-2.6	0.00931
ETHN_BL	0.291758	1.33878	0.064063	4.554	5.26E-06
ETHN_HM	0.858316	2.359185	0.33363	2.573	0.01009
ETHN_HO	0.875152	2.399239	0.360104	2.43	0.01509
ETHN_HP	0.675308	1.964637	0.596507	1.132	0.25759
ETHN_WH	0.050288	1.051573	0.06226	0.808	0.41926
ACT_English	-0.03331	0.967239	0.007746	-4.3	1.71E-05
ACT_Math	-0.014869	0.985241	0.008098	-1.836	0.06636
ACT_Reading	-0.018038	0.982124	0.007631	-2.364	0.01809
ACT_Science	-0.023999	0.976287	0.008328	-2.882	0.00396
ACT_Composit	0.073793	1.076584	0.025764	2.864	0.00418

dropped or failed credits and credit hours attempts. When we used those attributes along with pre-enrollment variables in proposed *TD-Cox* method, we get better classification performance. Thus, unlike other classification methods, the proposed *TD-Cox* approach has the ability to utilize extra semester-wise information by introducing time-dependent variables in the model.

Figure 4.5 and 4.6 provide the performance comparison between all the methods for each semester for different experimental setup. 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 in 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 thesis (*Cox* and *TD-Cox*) are a better choice for predicting student dropout. On the other hand, even if we rely only on the pre-enrollment attributes, *Cox* provides a better performance compared to other machine learning based classification methods. This suggests that Cox regression model would be a better choice for longitudinal data classification problem 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 handling censoring data are two specific features of longitudinal data that survival models such as Cox can efficiently handle. It is worth to mention that by comparing the result of 10-fold cross validation (Tables 4.3 and 4.5) and the test data for year 2009 (Tables 4.4 and 4.6) we can observe the clear benefits of using the proposed model.

4.5 Evaluation on Predicting Semester of Dropout

One of the primary purposes of this study is to build a model to estimate the semester of dropout using only the pre-enrollment information at the beginning of the study. As discussed earlier, one of the drawbacks of using linear regression in the presence of

Table 4.3: Performance of Logistic regression, Adaboost and Decision tree with Coxph 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.

	Model	**Accuracy**	**F-measure**	**AUC**
1st Semester	**Logistic**	0.705 (0.023)	0.702 (0.031)	0.734 (0.015)
	AdaBoost	0.709 (0.019)	0.712 (0.028)	0.747 (0.013)
	Decision tree	0.706 (0.035)	0.7 (0.049)	0.662 (0.026)
	Coxph	**0.719 (0.015)**	**0.724 (0.029)**	**0.751 (0.013)**
	TD-Coxph	**0.719 (0.015)**	**0.724 (0.029)**	**0.751 (0.013)**
2nd Semester	**Logistic**	0.715 (0.025)	0.715 (0.033)	0.766 (0.018)
	AdaBoost	0.721 (0.02)	0.724 (0.029)	0.78 (0.015)
	Decision tree	0.713 (0.037)	0.711 (0.051)	0.697 (0.023)
	Coxph	0.729 (0.019)	0.737 (0.031)	0.783 (0.012)
	TD-Coxph	**0.765 (0.018)**	**0.741 (0.03)**	**0.792 (0.011)**
3rd Semester	**Logistic**	0.728 (0.024)	0.729 (0.034)	0.795 (0.017)
	AdaBoost	0.733 (0.019)	0.734 (0.033)	0.802 (0.014)
	Decision tree	0.723 (0.034)	0.723 (0.048)	0.727 (0.019)
	Coxph	0.743 (0.018)	0.744 (0.028)	0.804 (0.013)
	TD-Coxph	**0.778 (0.016)**	**0.761 (0.029)**	**0.811 (0.012)**
4th Semester	**Logistic**	0.741 (0.021)	0.741 (0.032)	0.816 (0.016)
	AdaBoost	0.738 (0.027)	0.742 (0.03)	0.82 (0.015)
	Decision tree	0.727 (0.031)	0.734 (0.049)	0.738 (0.021)
	Coxph	0.747 (0.017)	0.747 (0.029)	0.831 (0.014)
	TD-Coxph	**0.801 (0.018)**	**0.784 (0.027)**	**0.835 (0.013)**
5th Semester	**Logistic**	0.748 (0.025)	0.746 (0.034)	0.824 (0.016)
	AdaBoost	0.751 (0.029)	0.751 (0.031)	0.826 (0.015)
	Decision tree	0.734 (0.039)	0.739 (0.039)	0.755 (0.019)
	Coxph	0.756 (0.02)	0.754 (0.027)	0.832 (0.012)
	TD-Coxph	**0.812 (0.019)**	**0.801 (0.026)**	**0.84 (0.012)**
6th Semester	**Logistic**	0.762 (0.028)	0.758 (0.03)	0.827 (0.014)
	AdaBoost	0.755 (0.023)	0.755 (0.035)	0.828 (0.016)
	Decision tree	0.745 (0.034)	0.742 (0.045)	0.75 (0.021)
	Coxph	0.767 (0.017)	0.767 (0.028)	0.836 (0.011)
	TD-Coxph	**0.821 (0.015)**	**0.818 (0.024)**	**0.847 (0.009)**

Table 4.4: Performance of Logistic regression, Adaboost and Decision tree with Coxph and TD-Cox on 2009 WSU student retention (experiment 1) for each semester along with standard deviation.

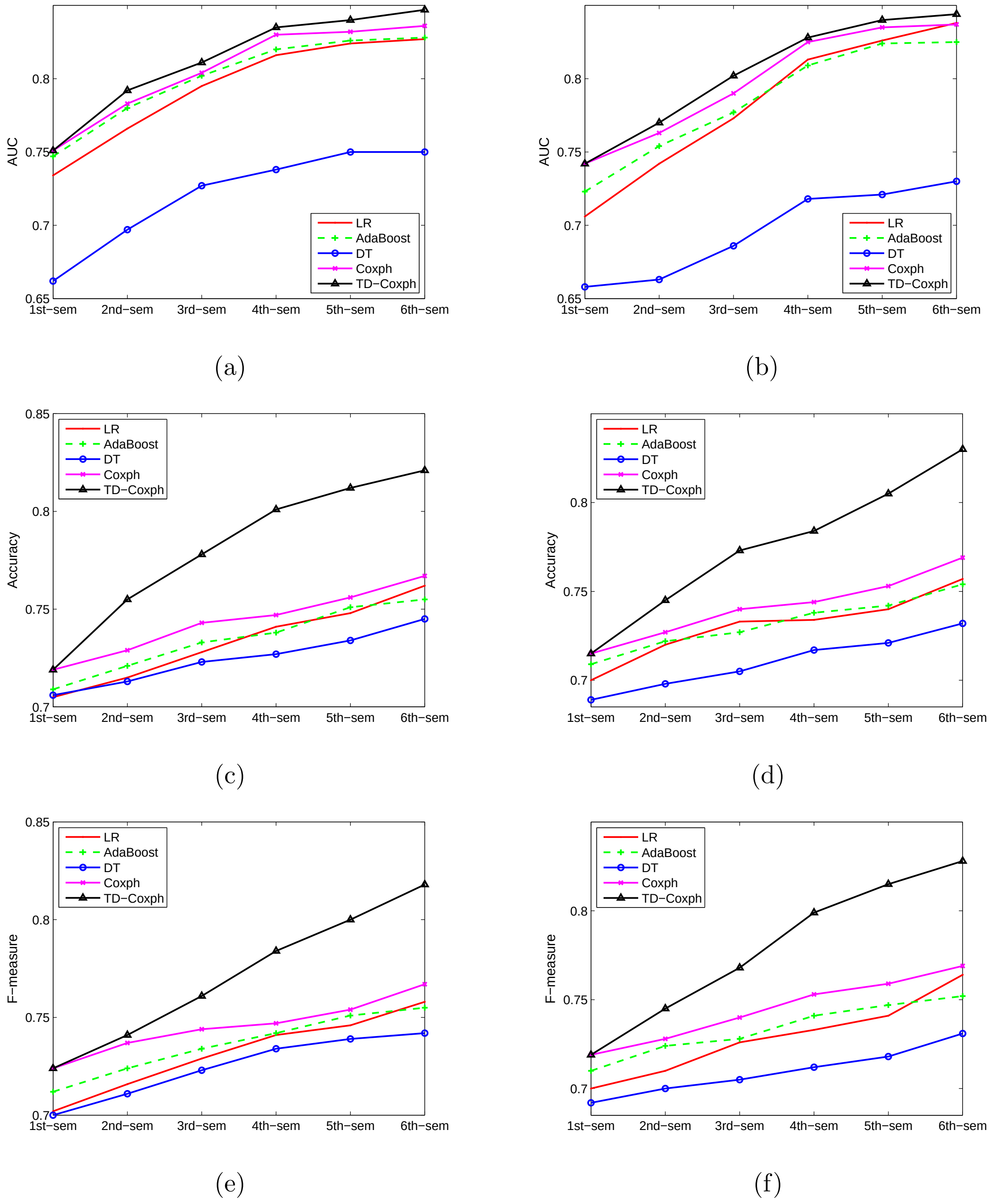
	Model	**Accuracy**	**F-measure**	**AUC**
1st Semester	**Logistic**	0.701 (0.019)	0.703 (0.025)	0.706 (0.016)
	AdaBoost	0.709 (0.017)	0.710 (0.027)	0.723 (0.012)
	Decision tree	0.689 (0.025)	0.692 (0.03)	0.658 (0.017)
	Coxph	**0.715 (0.018)**	**0.719 (0.024)**	**0.742 (0.014)**
	TD-Coxph	**0.715 (0.018)**	**0.719 (0.024)**	**0.742 (0.014)**
2nd Semester	**Logistic**	0.720 (0.018)	0.711 (0.029)	0.742 (0.015)
	AdaBoost	0.722 (0.019)	0.724 (0.026)	0.754 (0.013)
	Decision tree	0.698 (0.027)	0.701 (0.029)	0.663 (0.019)
	Coxph	0.727 (0.019)	0.728 (0.021)	0.763 (0.015)
	TD-Coxph	**0.745 (0.02)**	**0.745 (0.023)**	**0.77 (0.016)**
3rd Semester	**Logistic**	0.733 (0.019)	0.726 (0.027)	0.773 (0.015)
	AdaBoost	0.727 (0.02)	0.728 (0.025)	0.777 (0.013)
	Decision tree	0.705 (0.024)	0.705 (0.03)	0.686 (0.018)
	Coxph	0.74 (0.015)	0.740 (0.023)	0.790 (0.013)
	TD-Coxph	**0.773 (0.016)**	**0.768 (0.021)**	**0.802 (0.012)**
4th Semester	**Logistic**	0.734 (0.021)	0.733 (0.026)	0.813 (0.014)
	AdaBoost	0.738 (0.016)	0.741 (0.024)	0.809 (0.016)
	Decision tree	0.717 (0.023)	0.712 (0.033)	0.718 (0.017)
	Coxph	0.744 (0.017)	0.753 (0.023)	0.825 (0.01)
	TD-Coxph	**0.784 (0.014)**	**0.799 (0.024)**	**0.828 (0.014)**
5th Semester	**Logistic**	0.740 (0.018)	0.741 (0.025)	0.826 (0.012)
	AdaBoost	0.742 (0.019)	0.747 (0.028)	0.824 (0.014)
	Decision tree	0.721 (0.021)	0.718 (0.031)	0.721 (0.018)
	Coxph	0.753 (0.017)	0.759 (0.025)	0.835 (0.011)
	TD-Coxph	**0.805 (0.016)**	**0.815 (0.022)**	**0.84 (0.012)**
6th Semester	**Logistic**	0.757 (0.017)	0.764 (0.024)	0.838 (0.013)
	AdaBoost	0.754 (0.016)	0.752 (0.027)	0.825 (0.015)
	Decision tree	0.732 (0.02)	0.731 (0.029)	0.73 (0.017)
	Coxph	0.769 (0.014)	0.769 (0.022)	0.837 (0.009)
	TD-Coxph	**0.83 (0.013)**	**0.828 (0.02)**	**0.844 (0.009)**

Table 4.5: Performance of Logistic regression, Adaboost and Decision tree with Coxph 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.

	Model	**Accuracy**	**F-measure**	**AUC**
1st Semester	**Logistic**	0.705 (0.023)	0.702 (0.031)	0.734 (0.015)
	AdaBoost	0.709 (0.019)	0.712 (0.028)	0.747 (0.013)
	Decision tree	0.706 (0.035)	0.7 (0.049)	0.662 (0.026)
	Coxph	**0.719 (0.015)**	**0.724 (0.029)**	**0.751 (0.013)**
	TD-Coxph	**0.719 (0.015)**	**0.724 (0.029)**	**0.751 (0.013)**
2nd Semester	**Logistic**	0.715 (0.025)	0.715 (0.033)	0.766 (0.018)
	AdaBoost	0.721 (0.02)	0.724 (0.029)	0.78 (0.015)
	Decision tree	0.713 (0.037)	0.711 (0.051)	0.697 (0.023)
	Coxph	0.729 (0.019)	0.737 (0.031)	0.783 (0.012)
	TD-Coxph	**0.765 (0.018)**	**0.741 (0.03)**	**0.792 (0.011)**
3rd Semester	**Logistic**	0.718 (0.026)	0.724 (0.036)	0.779 (0.019)
	AdaBoost	0.724 (0.021)	0.729 (0.033)	0.788 (0.014)
	Decision tree	0.713 (0.036)	0.717 (0.049)	0.706 (0.02)
	Coxph	0.737 (0.019)	0.740 (0.029)	0.807 (0.015)
	TD-Coxph	**0.772 (0.017)**	**0.752 (0.029)**	**0.816 (0.014)**
4th Semester	**Logistic**	0.727 (0.023)	0.732 (0.033)	0.801 (0.016)
	AdaBoost	0.729 (0.029)	0.731 (0.034)	0.805 (0.017)
	Decision tree	0.718 (0.034)	0.720 (0.048)	0.715 (0.018)
	Coxph	0.742 (0.018)	0.745 (0.028)	0.826 (0.015)
	TD-Coxph	**0.79 (0.017)**	**0.774 (0.029)**	**0.830 (0.013)**
5th Semester	**Logistic**	0.730 (0.027)	0.736 (0.035)	0.807 (0.016)
	AdaBoost	0.732 (0.03)	0.739 (0.032)	0.810 (0.017)
	Decision tree	0.721 (0.039)	0.723 (0.041)	0.728 (0.018)
	Coxph	0.751 (0.021)	0.752 (0.027)	0.829 (0.015)
	TD-Coxph	**0.804 (0.019)**	**0.790 (0.027)**	**0.838 (0.014)**
16th Semester	**Logistic**	0.741 (0.031)	0.743 (0.033)	0.812 (0.014)
	AdaBoost	0.735 (0.025)	0.741 (0.035)	0.815 (0.015)
	Decision tree	0.729 (0.037)	0.730 (0.046)	0.741 (0.022)
	Coxph	0.760 (0.019)	0.760 (0.026)	0.832 (0.013)
	TD-Coxph	**0.817 (0.017)**	**0.811 (0.025)**	**0.840 (0.009)**

Table 4.6: Performance of Logistic regression, Adaboost and Decision tree with Coxph and TD-Cox on 2009 WSU student retention (experiment 2) for each semester along with standard deviation.

	Model	**Accuracy**	**F-measure**	**AUC**
1st Semester	**Logistic**	0.701 (0.019)	0.703 (0.025)	0.706 (0.016)
	AdaBoost	0.709 (0.017)	0.710 (0.027)	0.723 (0.012)
	Decision tree	0.689 (0.025)	0.692 (0.03)	0.658 (0.017)
	Coxph	**0.715 (0.018)**	**0.719 (0.024)**	**0.742 (0.014)**
	TD-Coxph	**0.715 (0.018)**	**0.719 (0.024)**	**0.742 (0.014)**
2nd Semester	**Logistic**	0.720 (0.018)	0.711 (0.029)	0.742 (0.015)
	AdaBoost	0.722 (0.019)	0.724 (0.026)	0.754 (0.013)
	Decision tree	0.698 (0.027)	0.701 (0.029)	0.663 (0.019)
	Coxph	0.727 (0.019)	0.728 (0.021)	0.763 (0.015)
	TD-Coxph	**0.745 (0.02)**	**0.745 (0.023)**	**0.77 (0.016)**
3rd Semester	**Logistic**	0.725 (0.022)	0.724 (0.027)	0.759 (0.015)
	AdaBoost	0.726 (0.023)	0.727 (0.025)	0.761 (0.014)
	Decision tree	0.702 (0.026)	0.703 (0.03)	0.684 (0.018)
	Coxph	0.733 (0.018)	0.735 (0.023)	0.784 (0.014)
	TD-Coxph	**0.765 (0.017)**	**0.760 (0.021)**	**0.800 (0.013)**
4th Semester	**Logistic**	0.727 (0.023)	0.734 (0.03)	0.782 (0.017)
	AdaBoost	0.729 (0.019)	0.735 (0.025)	0.774 (0.018)
	Decision tree	0.708 (0.024)	0.710 (0.034)	0.705 (0.019)
	Coxph	0.739 (0.02)	0.747 (0.025)	0.812 (0.015)
	TD-Coxph	**0.774 (0.016)**	**0.785 (0.024)**	**0.820 (0.016)**
5th Semester	**Logistic**	0.730 (0.019)	0.746 (0.027)	0.805 (0.015)
	AdaBoost	0.732 (0.018)	0.738 (0.029)	0.793 (0.016)
	Decision tree	0.712 (0.023)	0.712 (0.031)	0.715 (0.021)
	Coxph	0.748 (0.018)	0.756 (0.029)	0.822 (0.013)
	TD-Coxph	**0.792 (0.017)**	**0.811 (0.021)**	**0.835 (0.012)**
6th Semester	**Logistic**	0.740 (0.017)	0.750 (0.027)	0.815 (0.013)
	AdaBoost	0.739 (0.019)	0.741 (0.028)	0.810 (0.015)
	Decision tree	0.715 (0.021)	0.719 (0.029)	0.721 (0.019)
	Coxph	0.757 (0.015)	0.761 (0.024)	0.829 (0.012)
	TD-Coxph	**0.821 (0.013)**	**0.825 (0.021)**	**0.839 (0.01)**



(a)

(b)

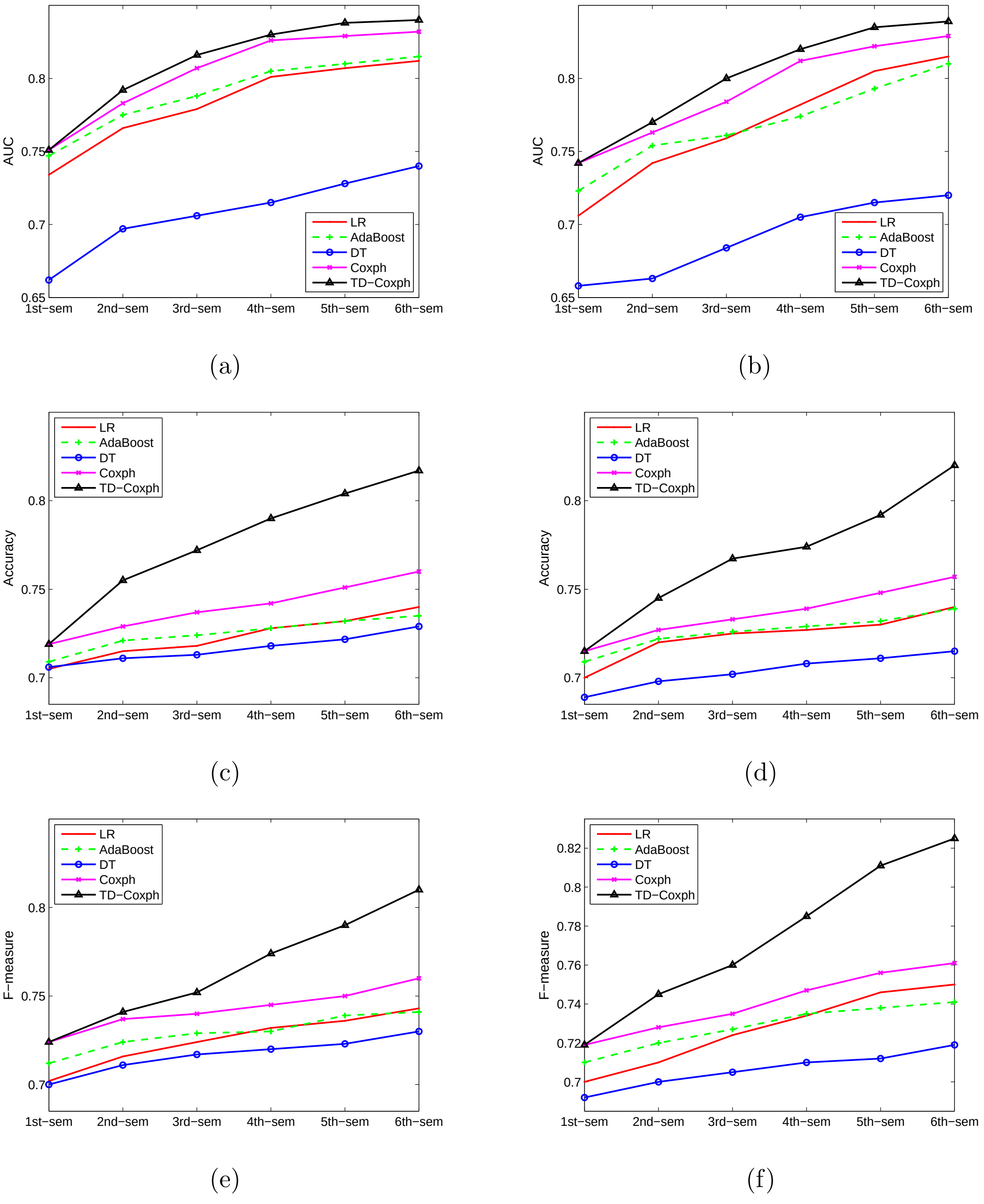
(c)

(d)

(e)

(f)

Figure 4.5: Performance of different methods obtained at different semesters (experiment 1): (a), (c) and (e) are the results of 10-fold cross validation on 2002-2008 training data and (b), (d) and (f) are the results for 2009 test dataset.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.6: Performance of different methods obtained at different semesters (experiment 2): (a), (c) and (e) are the results of 10-fold cross validation on 2002-2008 training data and (b), (d) and (f) are the results for 2009 test dataset.

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. Thus, regression methods cannot answer the important question of "when student is going to dropout". Therefore, in this thesis, we use *Cox* to answer this question. Tables 4.7 and 4.8 show the result of the 10-fold cross-validation training data and 2009 data as test data using first and second experimental setups, respectively. We compare the result of *Cox* with linear regression and well-known Support Vector Regression (SVR) [13].

Table 4.7: Performance of linear regression, SVR and Cox methods in predicting the semester of dropout on WSU student retention data (experiment 1) from 2002 to 2008 for each semester using 10-fold cross validation and 2009 student retention data.

	10-fold Cross Validation				**Test Data (Year 2009)**			
Model	**MAE**	**RMSE**	**RAE**	**RRSE**	**MAE**	**RMSE**	**RAE**	**RRSE**
Regression	1.79	1.96	0.746	0.752	1.83	1.99	0.734	0.737
SVR	1.83	2.14	0.769	0.826	1.92	2.17	0.826	0.847
Cox	**1.07**	**1.29**	**0.542**	**0.571**	**1.09**	**1.32**	**0.526**	**0.533**

Table 4.8: Performance of linear regression, SVR and Cox methods in predicting the semester of dropout on WSU student retention data (experiment 2) from 2002 to 2008 for each semester using 10-fold cross validation and 2009 student retention data.

	10-fold Cross Validation				**Test Data (Year 2009)**			
Model	**MAE**	**RMSE**	**RAE**	**RRSE**	**MAE**	**RMSE**	**RAE**	**RRSE**
Regression	1.91	2.07	0.763	0.772	1.98	2.11	0.734	0.737
SVR	1.97	2.38	0.791	0.835	2.04	2.43	0.838	0.871
Cox	**1.12**	**1.31**	**0.564**	**0.582**	**1.13**	**1.34**	**0.529**	**0.538**

It should also be mentioned 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 after the students start their semester. In other words, we are interested to estimate semester of dropout without using any semester-wise information. We can conclude that the *Cox* method provided more accurate estimation of the semester of dropout (using only pre-enrollment data) compared to the linear regression and SVR. This will allow us to have more focus towards specific high-risk student as early as (s)he starts the school. Comparing the results in Table 4.7 with Table 4.8, it is clear that in the presence of censored data, survival based methods such *Cox* have better performance compared to the traditional methods such as regression. Using this approach, we can accurately identify the student with higher risk of dropout and invest more effort on them, thus maximizing the retention rate which can then translate into increasing graduation from the university. Equally important, 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.

CHAPTER 5

CONCLUSION AND FUTURE WORK

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 implies the requirement of a longitudinal modeling approach. Benefits of survival analysis as an approach for studying the timing of events are clear in many different application domains which deal with longitudinal data. In this thesis, 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 during their early stage of higher educational life. In our work, extending survival analysis to the study of retention has provided an ability to study the temporal nature of the attrition behaviors.

Our study has shown the benefits of survival analysis as a methodology for the study of college student dropout behaviors. It should be noted that the majority of dropouts happen during freshman year (first two semesters). Thus, the ability to build a model that provides the prediction result at an early stage with high accuracy is very crucial. In this thesis, we took advantage of pre-enrollment information as well as 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. Once identified, these at-risk students can then be targeted with academic and administrative support to increase their chance of staying in the program.

The findings of the present study support the intuitively appealing conclusion that those students who have better semester GPAs are more likely to remain in school for the next semester. Motivated by this work at Wayne State University, the proposed method allows educational institutions to undertake timely measures and actions in their student attrition problem. Based on the findings of this thesis, we can use pre-enrollment information as screening test to identify students who are at a higher risk of dropping out of their study. It also shows that using the number of withdrawn or passed credits and GPA at each semester as an early warning to intervene when the students are doing poorly is critically important. 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 in a course to improve student success within the course, and in turn, reduce the student dropouts.

REFERENCES

[1] Alkhasawneh, R.: Developing a hybrid model to predict student first year retention and academic success in stem disciplines using neural networks. Ph.D. thesis, Virginia Commonwealth University (2011)

[2] Andersen, P., Gill, R.: Cox regression model for counting process: A large sample study. The Annals of Statistics **10**(4), 1100–1120 (1982)

[3] Bhardwaj, B.K., Pal, S.: Data mining: A prediction for performance improvement using classification. International Journal of Computer Science and Information Security **9**(4), 136–140 (2011)

[4] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92, pp. 144–152. ACM, New York, NY, USA (1992)

[5] Breslow, N.E.: Analysis of survival data under the proportional hazards model. International Statistical Review/Revue Internationale de Statistique pp. 45–57 (1975)

[6] Cox, D.: Some remarks on the analysis of survival data. In: Proceedings of the First Seattle Symposium in Biostatistics, pp. 1–9. Springer (1997)

[7] Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) pp. 187–220 (1972)

[8] DeBerard, M.S., Spielmans, G., Julka, D.: Predictors of academic achievement and retention among college freshmen: A longitudinal study. College student journal **38**(1), 66–80 (2004)

[9] Deike, R.C.: A study of college student graduation using discrete time survival analysis. Ph.D. thesis, The Pennsylvania State University (2003)

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

[11] Dey, E.L., Astin, A.W.: Statistical alternatives for studying college student reten-

tion: A comparative analysis of logit, probit, and linear regression. Research in Higher Education **34**(5), 569–581 (1993)

[12] Draper, N.R., Smith, H., Pownell, E.: Applied regression analysis, vol. 3. Wiley New York (1966)

[13] Drucker, H., Burges, C.J., Kaufman, L., Smola, A.J., Vapnik, V.: Support vector regression machines. In: Advances in Neural Information Processing Systems, pp. 155–161 (1997)

[14] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences **55**(1), 119–139 (1997)

[15] Glynn, J.G., Sauer, P.L., Miller, T.E.: A logistic regression model for the enhancement of student retention: The identification of at-risk freshmen. International Business & Economics Research Journal (IBER) **1**(8), 79–86 (2011)

[16] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter **11**(1), 10–18 (2009)

[17] Herzog, S.: 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)

[18] Hosmer, D.W., Lemeshow, S.: Applied survival analysis: regression modeling of time to event data. Wiley, New York (1999)

[19] Ishitani, T.T.: 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] Ishitani, T.T.: Studying attrition and degree completion behavior among first-

generation college students in the united states. Journal of Higher Education pp. 861–885 (2006)

[21] Ishitani, T.T., DesJardins, S.L.: 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] Jones-White, D.R., Radcliffe, P.M., Huesman Jr, R.L., Kellogg, J.P.: 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] Lee, E.T., Wang, J.: Statistical methods for survival data analysis, vol. 476. John Wiley & Sons (2003)

[24] Lin, J., Imbrie, P., Reid, K.J.: Student retention modelling: An evaluation of different methods and their impact on prediction results. Research in Engineering Education Sysmposium pp. 1–6 (2009)

[25] Luna, J.: Predicting student retention and academic success at new mexico tech. Ph.D. thesis, New Mexico Institute of Mining and Technology (2000)

[26] Miller, R.G., Halpern, J.: Regression with Censored Data. Biometrika Trust **69**(3), 521–531 (1982)

[27] Murtaugh, P.A., Burns, L.D., Schuster, J.: Predicting the retention of university students. Research in higher education **40**(3), 355–371 (1999)

[28] Nandeshwar, A., Menzies, T., Nelson, A.: Learning patterns of university student retention. Expert Systems with Applications **38**(12), 14,984–14,996 (2011)

[29] Pandey, U.K., Pal, S.: Data mining: A prediction of performer or underperformer using classification. International Journal of Computer Science and Information Technology **2**(2), 686–690 (2011)

[30] Pascarella, E.T., Terenzini, P.T., Feldman, K.A.: How college affects students, vol. 2. Jossey-Bass San Francisco, CA (2005)

[31] Peoples-McAfee, L.: A Longitudinal Analysis Using Auxiliary Information to Model Retention in Undergraduate Students. Texas Woman's University (2009)

[32] Quadri, M., Kalyankar, N.: 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)

[33] Quinlan, J.R.: C4. 5: programs for machine learning. Elsevier (2014)

[34] Radcliffe, P.M., Huesman Jr, R.L., Kellogg, J.P.: Identifying students at risk: Utilizing survival analysis to study student athlete attrition. IR Applications **12**, 1–10 (2009)

[35] Thammasiri, D., Delen, D., Meesad, P., Kasap, N.: A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications **41**(2), 321–330 (2014)

[36] Therneau, T.M., Lumley, T.: Package survival (2015). URL `https://cran.r-project.org/web/packages/survival/index.html`

[37] Thomas, L.: Student retention in higher education: the role of institutional habitus. Journal of Education Policy **17**(4), 423–442 (2002)

[38] Tinto, V.: Leaving college: Rethinking the causes and cures of student attrition. ERIC (1987)

[39] Tinto, V.: Research and practice of student retention: what next? Journal of College Student Retention: Research, Theory & Practice **8**(1), 1–19 (2006)

[40] Trevor, H., Robert, T., Jerome, F.: The elements of statistical learning: data mining, inference and prediction. New York: Springer-Verlag **1**(8), 371–406 (2001)

[41] Yadav, S.K., Bharadwaj, B., Pal, S.: Mining education data to predict student's

retention: A comparative study. International Journal of Computer Science and Information Security **10**(2), 113 (2012)

[42] Yu, C.H., DiGangi, S., Jannasch-Pennell, A., Kaprolet, C.: A data mining approach for identifying predictors of student retention from sophomore to junior year. Journal of Data Science **8**, 307–325 (2010)

[43] Zhang, G., Anderson, T.J., Ohland, M.W., Thorndyke, B.R.: Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. Journal of Engineering education **93**(4), 313–320 (2004)

[44] Zhang, Y., Oussena, S., Clark, T., Hyensook, K.: Using data mining to improve student retention in he: a case study. In: Proceedings of the 12th International Conference on Enterprise Information Systems (ICEIS 2010), pp. 190–197 (2010)

ABSTRACT

SURVIVAL ANALYSIS APPROACH FOR EARLY PREDICTION OF STUDENT DROPOUT

by

SATTAR AMERI

December 2015

Advisor: Dr. Chandan Reddy

Major: Computer Science

Degree: Master of Science

Retention of students at colleges and universities has long been a concern for educators for many decades. The consequences of student attrition are significant for both students, academic staffs and the overall institution. Thus, increasing student retention is a long term goal of any academic institution. The most vulnerable students at all institutions of higher education are the freshman students, who are at the highest risk of dropping out at the beginning of their study. Consequently, the early identification of *"at-risk"* students is a crucial task that needs to be addressed precisely. In this thesis, we develop a framework for early prediction of student success using survival analysis approach. We propose time-dependent Cox (TD-Cox), which is based on the Cox proportional hazard regression model and also captures time-varying factors to address the challenge of predicting dropout students as well as the semester that the dropout will occur, to enable proactive interventions. This is critical in student retention problem because not only correctly classifying whether student is going to dropout is important but also when this is going to happen is crucial to investigate. We evaluate our method on real student data collected at Wayne State University. The results show that the proposed Cox-based framework can predict the student dropout and the semester of dropout with high accuracy and precision compared to the other alternative state-of-the-art methods.

AUTOBIOGRAPHICAL STATEMENT

Sattar Ameri

Sattar Ameri entered to the Computer Science Master program at Wayne State University in 2013. Currently, he is a Ph.D student at the same department. His main areas of research are data mining, machine learning, survival analysis, pattern recognition and image processing.