Summer 7-31-2017

Prediction of Graduation Delay Based on Student Characterisitics and Performance

Tushar Ojha
University of New Mexico - Main Campus

Tushar Ojha

Candidate

Electrical and Computer Engineering

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Gregory L Heileman , Chairperson

Manel Martìnez-Ramòn

Don R Hush

Ahmad Slim

Prediction of Graduation Delay Based on Student Characteristics and Performance

by

Tushar Ojha

B.E., Birla Institute of Technology, 2015

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Engineering

The University of New Mexico

Albuquerque, New Mexico

July, 2017

Dedication

This thesis is dedicated to my parents.

Acknowledgments

I would like to thank my committee members, Dr. Gregory L. Heileman, Dr. Manel Martinez-Ramon, Dr. Don Hush and Dr. Ahmad Slim. I would sincerely like to thank Dr. Heileman for giving me an opportunity to learn and grow over the duration of my master's degree and for his invaluable guidance as my advisor. I would also like to extend heartfelt gratitude towards Dr. Martinez and Dr. Slim for their collaboration and invaluable inputs throughout the duration of this work. Much gratitude is also extended to Mr. Aman Sawhney and Mr. Divya Jyoti Prakash for their constant help and support. Last but not the least, I would like to thank everyone at IDI for making learning and working very enjoyable.

Prediction of Graduation Delay Based on Student Characteristics and Performance

by

Tushar Ojha

B.E., Birla Institute of Technology, 2015

M.S., Computer Engineering, University of New Mexico, 2017

Abstract

A college student's success depends on many factors including pre-university characteristics and university student support services. Student graduation rates are often used as an objective metric to measure institutional effectiveness. This work studies the impact of such factors on graduation rates, with a particular focus on delay in graduation. In this work, we used feature selection methods to identify a subset of the pre-institutional features with the highest discriminative power. In particular, Forward Selection with Linear Regression, Backward Elimination with Linear Regression, and Lasso Regression were applied. The feature sets were selected in a multivariate fashion. High school GPA, ACT scores, student's high school, financial aid received, and first generation status were found to be important for predicting success. In order to predict delay in graduation, we trained predictive models using Support Vector Machines (SVMs), Gaussian Processes (GPs), and Deep Boltzmann Machines (DBMs) on real student data. The difference in performance among the models is negligible with respect to overall accuracies obtained. Further analysis showed that DBMs outperform SVMs in terms of precision and recall for individual

classes. However, the DBM and SVM implementations are computationally inexpensive compared to GPs, given the same resources.

Contents

List of Figures

List of Figures

List of Tables

Glossary

DBM Deep Boltzmann Machine.

Feature Quantitative or Qualitative property of an entity or event.

GPC Gaussian Process Classification.

SVM Support Vector Machine.

LASSO Least Absolute Shrinkage and Selection Operator.

Multi-class The input is to be classified into one, and only one, of the n non-overlapping classes.

Precision It is the fraction of events where we correctly declare i out of all instances where the algorithm declared i.

Recall It is the fraction of events where we correctly declared i out of all the cases where the true state of the world is i.

F-Score Harmonic average of precision and recall.

Error rate Percent of misclassified records out of the total records in the validation data.

GPA Grade Point Average.

HS High School.

HSGPA The high school GPA of student at the time of graduation.

ACT American College Testing.

SAT Scholastic Aptitude Test.

Learning Learning can be viewed as a process where the input consists of any

available data, from which information is extracted, and then knowledge is inferred from this information.

Chapter 1

Introduction

1.1 Overview

University education aims to impart students with the necessary skills in their chosen vocation. Its value is undisputed. A natural inclination, therefore, is to judge its effectiveness. How effective is a university's methodology in nurturing productive student growth? As part of a defined curriculum, each university lays out a system to grade how well students gain the desired skills. Although most curricula are designed for a four year span, in mind the average six year graduation rate in US is a mere 60% [35]. This makes the problem of determining of a university's effectiveness much more pertinent.

For a large public university, such as the University of New Mexico (UNM), the task is even more daunting. Large universities are a vast network of interconnected systems, and therefore singling out a bottleneck is extremely difficult. The ultimate measure of success is a university's graduation rate [56, 49, 48, 47, 50]. Universities proactively try to predict graduation rates. Most studies use this statistic to qualitatively assess the factors influencing student progress.

This work leverages student data to predict graduation delay for a student and student cohorts (students belonging to the same incoming class). It provides a means to study the factors that influence the amount of time it takes for a student to graduate. An essential set of the factors that influence graduation rates is constituted from student characteristics, i.e. his/her previous record, socio-economic background, etc. To efficiently roll out initiatives aimed at student success, there is a need to study, analyze and quantify the factors that have a direct impact on student success. The challenge is to identify these factors, and engineer solutions based on them, to produce quantifiable results that could be used to improve student progress at universities.

Graduation rates are affected by a number of factors that can be grouped under two categories: pre-institutional factors and institutional factors [25, 55]. The former includes factors such as race, ethnicity, high school preparation and socio-economic status, while the latter includes factors such as tutoring, advisement arrangements, competence of instructors and curricular complexity [49, 48]. This is illustrated in Figure 1.1.

We begin our work by identifying a subset of features that best predict graduation rates. We employ various multivariate feature selection techniques to select an optimal subset of features to be used for classification. These selected features are then used to perform a multi-class classification using three predictive models, namely DBM, SVM and GPC, with the objective of predicting a delay in graduation.

The rest of this thesis is organized as follows: The remainder of this chapter considers the motivation behind this work and common student characteristics that are most widely attributed to be the factors that contribute to a student's success in a university. Chapter 2 describes in detail the feature selection techniques we used to find the best subset of features. Chapter 3 discusses the classification algorithms/-models we used for predicting the delay in student graduation. The procedure used

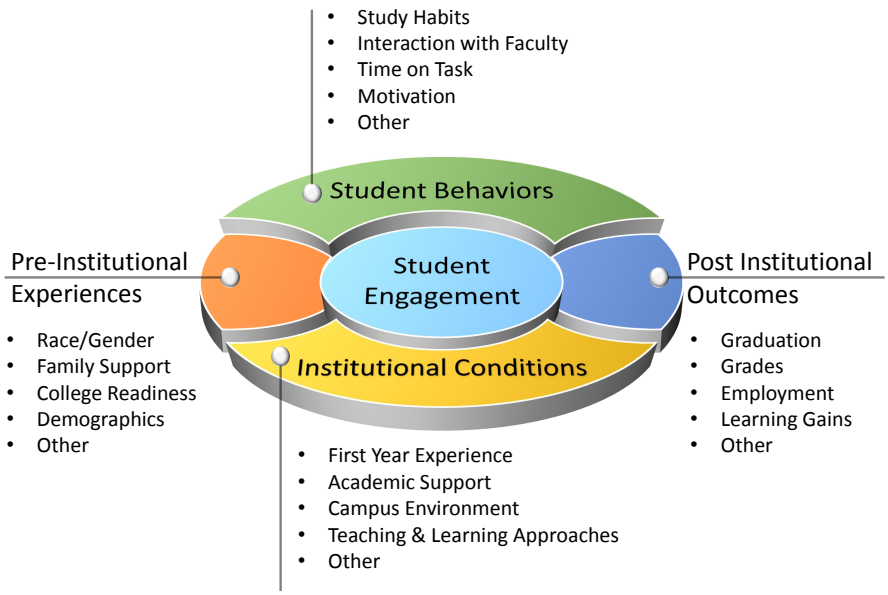

Figure 1.1: Student success framework.

for validation of results is discussed at length. The results are discussed using precision recall metrics for individual classes. Finally, concluding remarks are provided in the Chapter 4.

1.2 Motivation

According to the report "Performance-Based Funding for Higher Education" [38], thirty two states in the United States have adopted performance-based funding for four-year colleges and many other states are in the process of transitioning to this approach [38]. Some performance indicators used by these states include course completion, time to degree, transfer rates, the number of degrees awarded, and/or the number of low-income and minority graduates. These performance indicators influence institutional ranking and the funding it receives to some extent [1]. Cost of education is another factor that beckons the need for improving student success

metrics. The tuition at four-year colleges has more than doubled over the past three decades [37] and a delay in graduation results in an increase in debt incurred by the student.

In a span of 20 years, from 1992 to 2012, the average debt incurred by a student loan borrower who graduated with a bachelor's degree more than doubled to a total of nearly $27,000 [42]. For instance, according to the statistics reported by The University of Texas at San Antonio (UTSA), the average undergraduate student loan debt incurred by students who graduated in four years was $19,239. The amount borrowed increases sharply if students graduating are delayed by two years. The average debt in this case grows to $26,191. Figure 1.2 represents the average debt incurred by undergraduate students at UTSA who graduated in 2012-13. Four-year graduates began as freshmen in 2009-10, while five-year graduates started in 2008-09 and six-year graduates started in 2007-08.

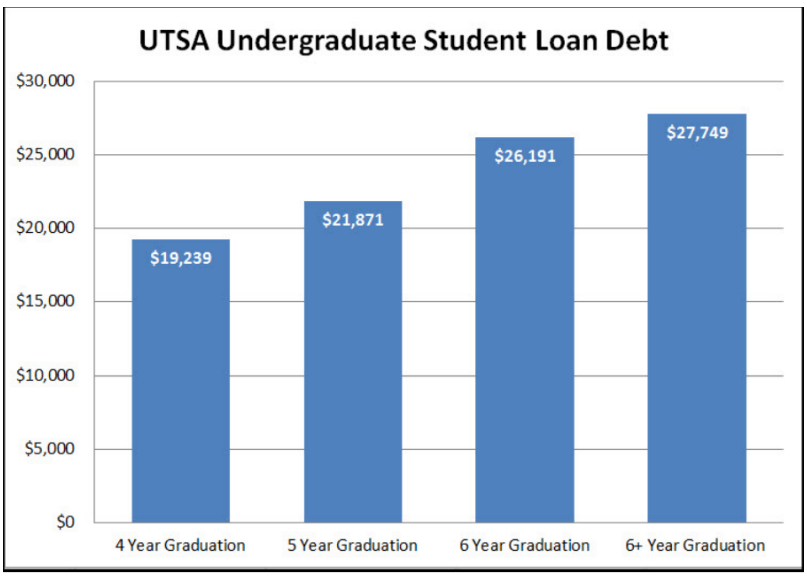

Figure 1.2: Loan indebtedness for 4 years versus 5 or 6 or 6+ years at UTSA [32].

Today, more than 40% of students who start as freshmen at four-year colleges do not graduate within six years [35]. The student borrowers who default have a median

debt of around $8,900 and an average debt of $14,500 [54]. In fact, higher graduation rates at four-year colleges result in fewer defaults, according to a United States Department of Education study [13]. This is illustrated in Figure 1.3. Therefore, it is imperative to focus on student success at universities to help students mitigate financial burdens before joining the workforce.

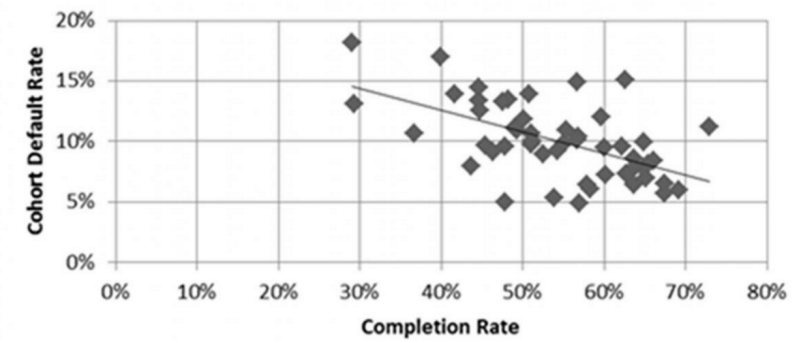

Figure 1.3: Higher graduation rates at four-year colleges lowers the default rates [13].

Another statistic that adds to this financial difficulty is the decreasing federal and state funding at four-year public institutions over the past decade. The federal and state funding per Full Time Equivalent (FTE) student at public four-year colleges in the 2003-04 academic year was $7,170 and $8,980 respectively. These numbers reduced to $6,910 and $7,110 for federal and state funding respectively. This was compensated by increases in tuition per FTE student from $6,610 in the 2003-04 academic year to $9,740 in the 2013-14 academic year [10]. Such changes increase student's burden, particularly if they fail to complete their degree or if they fail to graduate within six years.

The market requirement of a skilled workforce is growing, with more than 50% of upcoming jobs requiring a bachelor's degree or higher. It is estimated that by 2020, around two thirds of all job openings will require postsecondary education or training [9]. This provides an incentive for universities to graduate a higher number of competent workers. Thus, improving student success has become a critical task at

many institutions [39]. In addition, compared to their counterparts, college graduates earn 66% more and have a lower probability of unemployment [36]. Over the course of a lifetime, an average worker with a bachelor's degree will earn $1 million more than a worker without a postsecondary degree [8].

The profile of incoming students is increasingly becoming nontraditional. Many students no longer enroll in college straight after high school or live on campus and attend classes full time [16]. Instead, nearly one third of students attend class part time and around 26% work full time while attending classes. At the present time, 28% of incoming students take care of dependents while enrolled in college and 44% of the college and university students are 24 years old or older. Furthermore, 18% are non-native English speakers and about 42% come from families of color [16]. Thus, it is essential to identify these changing student demographics and characteristics and understand their effect on graduation rates.

1.3 Pre-institutional Factors

This section discusses the most commonly studied and cited pre-institutional student characteristics in literature that contribute to a student's success at postsecondary institutions. These factors include both educational and socio-economic student background.

1.3.1 High School GPA and SAT scores

Tables 1.1 and 1.2 show the results reported by the Cooperative Institutional Research Program (CIRP) freshman survey for the entering cohorts of 1994 and 2004. There is a monotonically increasing relationship between degree attainment and high school GPA and SAT scores, respectively. In particular, Table 1.1 shows that stu-

	% of Students holding Bachelor's degrees Within		
HSGPA	4 Years	5 Years	6 Years
A/A+	58.2	75.6	79.3
A-	47.8	66.3	70.6
B+	35.9	54.7	59.8
B	25.2	43.3	48.7
B-	15.5	30.5	36.6
C+	9.8	22.4	27.7
C or less	6.3	16.0	21.2

Table 1.1: Four, five and six year graduation rates by high school GPA.

dents with higher high school GPAs tend to graduate sooner than those with lower high school GPAs. This is also true for students with higher SAT scores.

1.3.2 Gender

Statistics show that, on average, women tend to graduate earlier than men. In the United States, for example, degree attainment rates for both genders have witnessed remarkable fluctuations through the years. Figure 1.4 shows that men used to have higher college degree attainment compared to women up until 2014 [46]. From 2013 back to 1967, the gap in degree attainment between men and women who are 25 years old and older ranged between 1% and 8% with a peak in 1983. In 2013 the gap went down to 1% with degree attainment at approximately 30% for the two genders. In 2015, the picture changed. At that time 33% of women 25 years old and older held a bachelor's degree or higher compared to 32% of men. This increase in degree attainment is driven by the increased involvement of women in higher education.

	% of Students holding Bachelor's degrees Within		
SAT score	4 Years	5 Years	6 Years
1300+	62.2	78.2	81.6
1200—1299	51.9	69.5	73.3
1100—1199	42.9	61.2	65.6
1000—1099	34.8	53.7	58.6
900—999	24.6	44.0	49.9
800-899	17.2	34.1	40.5
Less than 800	10.5	23.9	30.4

Table 1.2: Four, five and six year graduation rate by SAT score.

1.3.3 Ethnicity

Today, incoming first year students in the United States are more diverse than they were in the past, with a higher number of Hispanic, Black, part-time, older, low income and other minorities entering college. Furthermore, the graduation rates for these populations lag behind the more well-to-do white population [2]. According to the United States Census Bureau population survey statistics, Asian and non-Hispanic white ethnic groups have a relatively higher percentage of population 25 years old and older with a bachelor's degree. Although, over the years, the percentage of the population 25 years old or older having a bachelor's degree across all ethnicities has increased, the relative difference in percentages between these groups has remained the same, illustrated in the Figure 1.5. As can be seen in Figure 1.5, in 2015 the percentage of Black and Hispanic adults having a bachelor's degree is under 20%, whereas that for Asians is 54% and that for non-Hispanic whites is 36% [46].

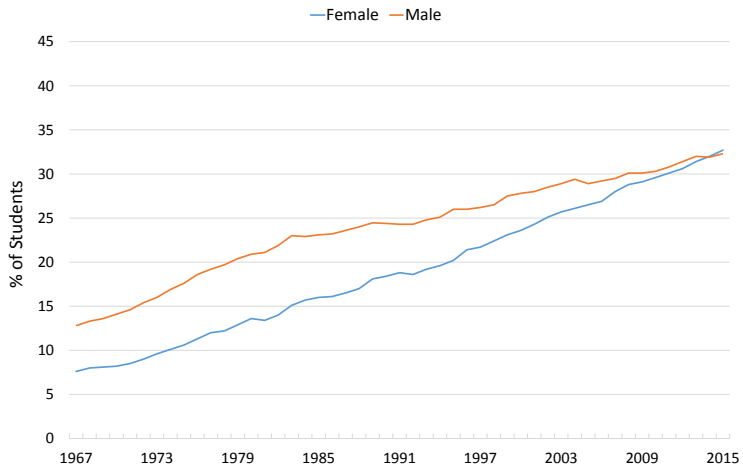

Figure 1.4: Percentage of the population 25 years old and older with a bachelor's degree or higher by gender from 1967 to 2015.

1.3.4 Financial

Students from the lowest quartile of family income have a very low bachelor's degree attainment rate as compared to those from the top quartile. The percentage of high school graduates in the lowest quartile continuing to college in 2015 was 62%, whereas it was 89% for the top quartile. If the continuation range for college in the year 2015 is considered, it is even worse, with the lowest quartile students at 45% as compared to top quartile students at 82%. Finally, the percentage gap increases even further when considering the eventual bachelor's degree graduation rates. Just 9% of students from the lowest quartile earn a bachelors degree by age 24, compared to 77% for the top quartile [41]. These statistics show family financial background to be a significant contributor towards bachelor's degree attainment.

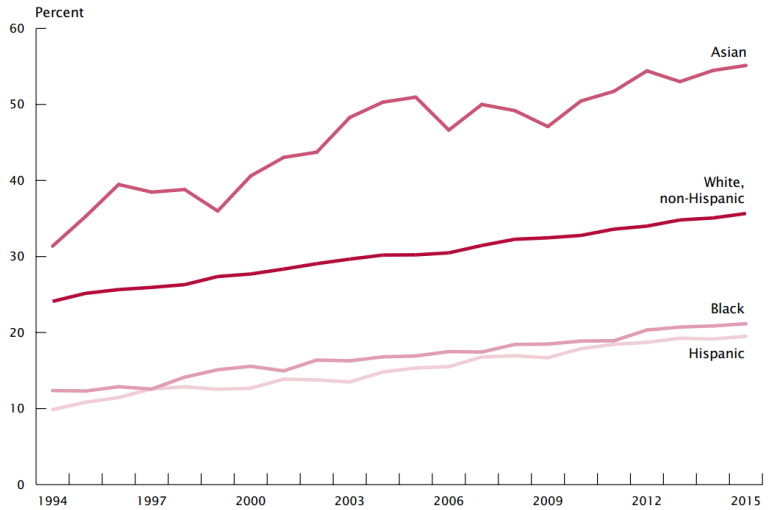

Figure 1.5: Percentage of the population 25 years old and older with a bachelor's degree or higher by race and hispanic origin from 1988 to 2015 [46].

1.3.5 Family Educational Background

Numerous studies have suggested that first-generation college students are at higher risk of dropping out of college than their non first-generation counterparts [58]. The percentage of first-generation students who earn a degree after four years is 27.4%, which is much less when compared to 42.1% for students who come from families with parents who have higher education experience [11]. This percentage gap remains consistent even when 6 year graduation rates are considered, as illustrated in Figure 1.6.

1.4 Institutional Factors

The first-year student characteristics measured in terms of GPA, number of credit hours and major changes can be very useful features in predicting whether or not a student eventually graduates [3]. A university typically requires a certain GPA to

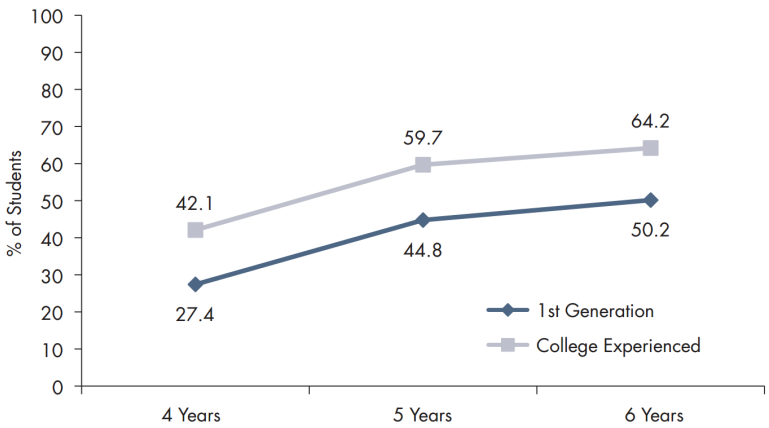

Figure 1.6: Weighted four, five and six year graduation rates, by generation in college [11].

be maintained and a certain number of credit hours to be completed as a condition for graduating. This implies that students GPA and number of credit hours are correlated with the eventual timely or delayed graduation of a student.

Chapter 2

Feature Selection

2.1 Introduction

Examining and understanding the intrinsic characteristics of the data is generally the starting point for any machine learning application. Feature selection involves choosing the best feature subset from a given dataset based on statistical and discriminative properties. Hence, it provides for a better quality feature set, i.e., less redundant and noisy, to be provided as input to learning algorithms. Feature selection is particularly useful when there are a large number of features. Most learning algorithm implementations require that all features be real-valued. However, in most real world scenarios the data is both categorical and real-valued. A direct translation of a categorical value set to real values is impractical, since real numbers have a natural numerical ordering (e.g. $2 > 1$ whereas $B \not> A$). A method commonly employed to solve this issue is to *binarize* the set of discrete (categorical) values, i.e., to map each categorical value to a binary number and treat each digit in this mapped space as a separate feature. Consequently, this leads to an expansion in the feature space.

One of the most widely used procedures to improve the quality of the feature set,

i.e., to reduce redundancy and noise in the feature set, is dimensionality reduction [52]. There are several techniques for dimensionality reduction in literature that can be broadly classified as feature extraction and feature selection. Feature extraction combines features and projects them into a new feature space with a reduction in the number of dimensions. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Canonical Correlation Analysis (CCA) [52] are a few prominent feature extraction techniques. Feature subset selection, according to Hall et al. [19], is the process of identifying and removing irrelevant and redundant information by selecting a subset of features in order to maximize relevance to the target, such as class labels [52]. It is also referred to as variable selection, variable subset selection, and attribute selection in machine learning and statistics literature.

Feature extraction and selection help to improve learning performance, lower computational complexity, build better generalizable models, and decrease required storage. Since the dimensions of the feature space are altered after feature extraction, interpreting features in terms of the original space is difficult [52]. This, however, is not the case while using feature selection, wherein the original dimensional space is maintained. Thus, feature selection edges out feature extraction in terms of interpretability [52], but not necessarily in terms of performance. Feature extraction (transformation) methods can be converted into feature selection methods via sparse learning techniques such as l_1 regularization [33].

In this work, we use real student data, which contains numerous categorical features, and binarizing them expands the feature space. Hence, we use feature selection on pre-institutional characteristics of university students, in order to improve the computational efficiency of our model.

2.2 Characteristics of Feature Selection

Typically there are four steps in feature selection: subset generation, subset evaluation, stopping criterion, and result validation [29]. The actions taken in each of these steps is discussed below:

1. **Subset Generation:** A search strategy is employed here to select the best feature subset using some evaluation criterion.
2. **Subset Evaluation:** In this step the set of features selected in the previous step are evaluated based on some evaluation criterion.
3. **Stopping Criterion:** Amongst all the selected subsets, the subset that best fits the evaluation criterion after the stopping criterion is met, is chosen.
4. **Result Validation** The selected subset is validated [52].

These processes are illustrated in Figure 2.1.

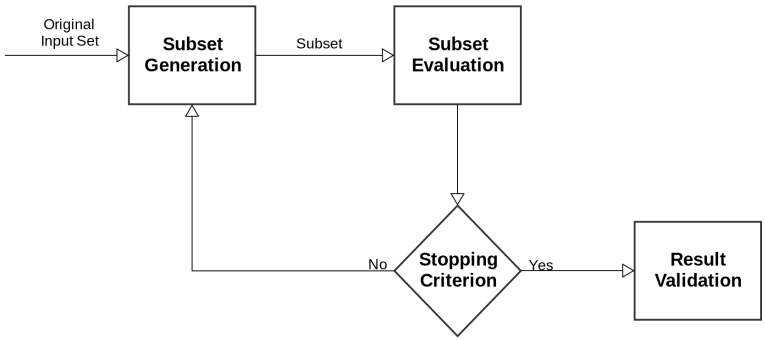

Figure 2.1: Feature selection steps.

Subset generation is a heuristic search in which a potential subset is defined by each state for evaluation in the search space. The subset search can be started with one feature in the set, with a feature added to the set at each step, which is called

forward selection. Whereas, if the set is initialized with the complete feature set and features are sequentially removed one at a time, it is termed as backward elimination. Naturally, the selection of a starting point is essential to this process. The subset generation process is characterized by successor generation and search organization [26]. The search starting point is determined in the successor generation phase, which in turn influences the search direction. To determine the search starting point at each state, forward, backward, compound, weighting, and random methods may be considered [12]. Search organization is responsible for the feature selection process with specific strategies like sequential search, exponential search [34] [40] or random search [28]. The evaluation criterion can be classified as independent or dependent, based on whether or not it uses the classification or regression models [29]. The independent criterion uses intrinsic characteristics of the training data to evaluate the importance of a feature or feature set, whereas the dependent criterion uses the training model for feature or feature set evaluation.

A term often used in feature selection literature is feature *relevance*. A feature X_i is considered to be strongly relevant if it creates a change in the probability distribution of the class values when used with the full feature set [19]. On the other hand, if X_i is not strongly relevant and creates a change in the probability distribution of the class values when used alongside a subset of the complete feature set, it termed as weakly relevant. All other features are considered as *irrelevant*.

2.3 Classification of Feature Selection Algorithms

The feature selection algorithms can be classified as supervised [57, 51], semi-supervised [60, 59] or unsupervised [15, 31].

Supervised feature selection algorithms can be subclassified as filter methods, wrapper methods [24], and embedded methods. A *filter method* uses heuristics based

on the intrinsic characteristics of the feature set, such as distance, consistency, dependency, or correlation, without the use of any learning algorithm [18]. This prevents against any bias that the learning algorithm might introduce into the feature selection process [52]. A *wrapper method* employs a predetermined learning algorithm to evaluate the feature set to be selected. Another class of methods, called *embedded methods* combine feature search and the learning algorithm into one optimization problem formulation.

Feature selection via wrapper and embedded methods are specific to a classifier [33]. Filters, on the other hand, totally ignore the effect of a selected feature subset on the performance of the induction algorithm [24, 19]. Since wrappers are tuned to the specific interaction between an induction algorithm and its training data, they generally outperform the filters [19]. However, wrappers are computationally intensive. Embedded methods, being a combination of filters and wrappers, are still relatively time efficient. In this work, we use wrapper and embedded methods for multivariate feature selection from the pre-institutional features dataset.

2.4 Wrapper Method

The wrapper approach employs, as a subroutine, a statistical re-sampling technique (such as cross-validation) using the actual target learning algorithm to estimate the accuracy of the feature subsets [24]. This use of the target learning algorithm comes from the assumption that the optimal feature subset should have a dependency on the specific biases and heuristics of the induction algorithm [52]. A generalized wrapper method performs the following steps, given a predefined learning algorithm:

1. **Search** This step involves searching for a subset of features.

2. **Evaluation** In this step, the subset of features selected from the previous step

is evaluated by the performance of the learning model.

3. **Stopping** The previous two steps are repeated until a desired quality is reached.

Figure 2.2 illustrates the general framework of feature selection. It involves three parts: Feature Search, Feature Evaluation, and Induction Algorithm.

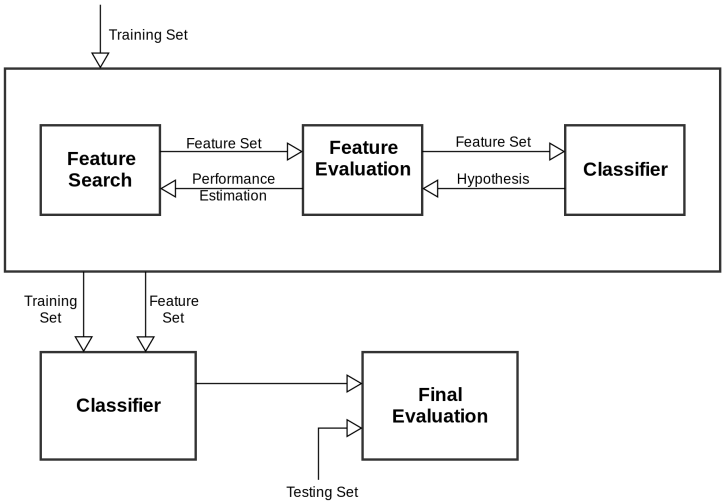

Figure 2.2: Generalized wrapper approach for feature selection.

Treating the learning algorithm as a black box, the feature search phase produces a set of features, which are fed to the feature evaluation phase. The evaluation phase uses a learning model to assess the performance which is sent back to the feature search phase to perform the next iteration of feature subset selection. This is performed until a feature subset with the best performance metric is found. The resulting learning model is then evaluated on a new validation set.

Suppose the feature space is of size m. Then the search space is of order $\mathcal{O}(2^m)$. If m is big, this exhaustive search is not practical. Instead, other search strategies like hill-climbing, best-first, branch and bound, and genetic algorithms can be used [17]. Table 2.1 lists the different search strategies. In this work, we use the greedy search

Algorithm Group	**Search Algorithm Name**
Exponential	Exhaustive search
	Branch-and-Bound
Sequential	Greedy forward selection or backward elimination
	Best-first
	Linear forward selection
	Floating forward or backward selection
	Beam search (and beam stack search)
	Race search
Randomized	Random generation
	Simulated annealing
	Evolutionary (e.g. genetic, colony etc.)
	Scatter search

Table 2.1: Search strategies for feature selection.

strategies with linear regression as the predefined learning model for the wrapper. Greedy search strategies are of two types: forward selection and backward elimination [52].

2.4.1 Forward Selection

Forward selection starts with an empty set of features. Features are constantly added to form an increasing set at each iteration based on a selection metric, until a stopping criterion is met.

2.4.2 Backward Elimination

As opposed to forward selection, backward elimination starts with a full set of features. In each iteration a feature is eliminated based on an elimination metric until a stopping criterion is met.

2.5 Embedded Method

As a first step, the embedded approach uses a statistical criterion to select numerous candidate feature subsets with a given cardinality (similar to filter methods). Thereon, the subset with the best evaluation metric value for the learning model is chosen [29]. Embedded models leverage the computational efficiency of filters and add sophistication using learning models as in wrappers. The embedded methods can be classified into the following three types [52]:

1. **Pruning Methods** At first, the model is trained using the complete feature set. Thereon, some features are zero weighted in an attempt to eliminate them, while trying to maintain model performance.

2. **Built-in Methods** Algorithms such as ID3 [45] and C4.5 [44] have a built-in mechanism for feature selection.

3. **Regularization Models** The regularization models use an objective function to minimize fitting errors, while reducing the feature weights to be very small, nearly or exactly zero. Thereafter, the features with zero weights are eliminated [30].

We use a regularization model in our feature selection process. Without loss of generality, we consider only linear models $\mathbf{w}$, wherein classification of $\mathbf{Y}$ can be based on a linear combination of $\mathbf{X}$ [52]. Here, $\mathbf{Y}$ denotes the vector of target values and $\mathbf{X}$ denotes the input feature vector. Here, $\mathbf{w}$ (weight) is estimated with properly tuned penalties. Each weight $w_i \in \mathbf{w}$, corresponds to a feature, f_i in the feature set. Only the features with a corresponding $w_i \neq 0$ are selected.
Mathematically, $\hat{\mathbf{w}}$ is given as:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, c(\mathbf{w}, \mathbf{X}) + \alpha \cdot penalty(\mathbf{w}) \tag{2.1}$$

where $c(.)$ is the classification objective function, $penalty(\mathbf{w})$ is a regularization term, and α is the regularization parameter controlling the trade-off between the $c(.)$ and the penalty. Quadratic loss (e.g. least squares), hinge loss (e.g. l_1SVM), and logistic loss are some of the most commonly used classification objective functions. For our experimentation we used the lasso regularization.

2.5.1 Lasso

Lasso regularization [53] is based on l_1 regularization of the coefficient $\mathbf{w}$ and defined as:

$$penalty(\mathbf{w}) = \sum_{i=1}^{m} |\mathbf{w}_i| \tag{2.2}$$

l_1 regularization has the property that it can generate an estimation of $\mathbf{w}$ with exact zero weights. These zero weights correspond to individual features which can be eliminated.

2.6 Experimental Setup

2.6.1 Dataset Description

The dataset used in our experiments is anonymized real student data from the University of New Mexico. It represents information for nearly two thousand First-Time Full-Time (FTFT) undergraduate students. Table 2.2 lists pre-institutional student characteristics that are present in the data.

Pre-Institutional Features	
High School GPA	HS GPA on "good" units
Enhanced ACT Reading	Passed HS Units Requirement
Enhance ACT Science Reasoning	HS Natural Science units
HS Social Science units	High School State: ACT Record
Enhanced ACT Composite	HS Math units
Age at Matriculation	HS Lab Science units
High School: ACT Record	HS History units
Gender	HS Foreign Language units
Ethnicity	HS total English units
Recent HS Graduate	HS units of English Composition
Current Residency Status (on-campus/off-campus)	Started in Summer Semester
HS Lack of English	HS Lack of Math
HS Lack of Foreign Language	HS Lack of Natural Science
HS Lack of Social Science	Original Place of Residence
Pell Grant Received	Pell Grant Eligible
First in family to go to University (First Generation)	

Table 2.2: Pre-institutional features used in feature selection.

The feature set is a mix of categorical and continuous valued features. Student's high school state, original place of residence, and ethnicity are a few examples of categorical features. During the preprocessing phase, categorical features were binarized and were thus, transformed into multiple binary valued features, based on the highest number of digits in the binarized space. After binarization, the total feature space expanded to a set of 56 features from a set of 31 features. Continuous features,

such as HSGPA, ACT scores, etc. were normalized as follows:

$$\frac{x - \mu}{\sigma}$$

Where x represents the set of values for a feature f, μ is the mean of the values of f, and σ is the standard deviation of f. This transformation of the values scales them to unit variance.

For imputation of missing values, we tried three different approaches which are as follows: replaced the null values of a feature with the mean of that feature, replaced the null values of a feature with the median of that feature, and omitted the records that had null values in any of the features. All the three approaches were marginally different. The results presented in Section 2.7 correspond to the imputation approach wherein the records having null value for any of the features were omitted.

2.6.2 Feature Selection Methods

We used three feature selection methods, namely forward selection with linear regression, backward elimination with linear regression, and lasso regression. The first two techniques are wrapper approaches, while the third technique is an embedded approach.

In the linear regression with forward selection and backward elimination, the R-squared value was used as the discriminative measure for features to be added to, or removed from the selected subset for the case of forward selection, or backward elimination, respectively. For all the three methods, the best feature subset for a given number of features was assessed using cross-validation mean squared error (CV-MSE). Additionally, the number of features in the selected subset was varied from zero to total number of features, i.e. 56. The performance for each number of features in the subset was gauged using CV-MSE.

2.7 Results and Discussion

The results for the three feature selection methods are illustrated in Figure 2.3. It is evident from Figure 2.3, while varying the number of features to be selected in the final subset, forward selection and backward elimination perform very similarly in terms of the CV-MSE. However, lasso regression constantly selects a smaller sized feature subset. Thus, it can be said that the forward selection and backward elimination outperform lasso regression.

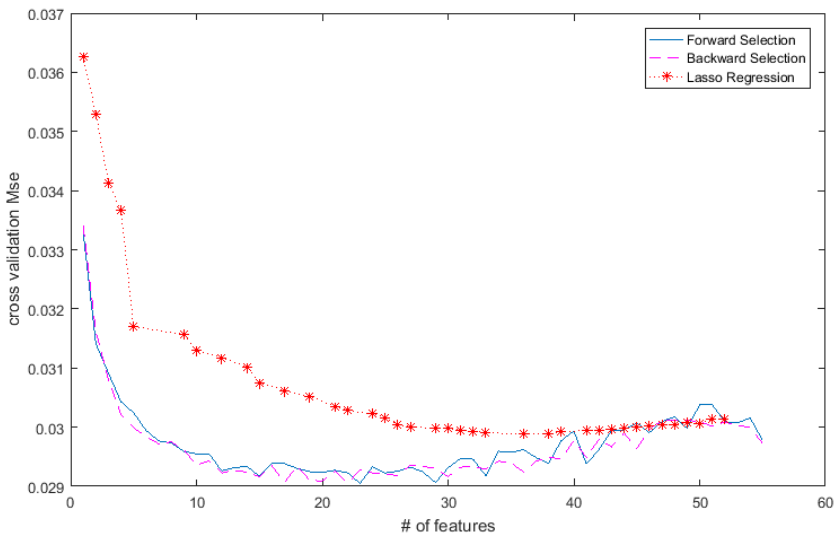

Figure 2.3: Number of features in the selected subset vs. the cross-validation mean squared error for forward selection, backward elimination, and lasso regression.

The minimum CV-MSE obtained with linear regression methods was 0.029 and using lasso regression yielded a minimum CV-MSE of 0.030. Additionally, the num-

Forward Selection	**Backward Elimination**	**Lasso**
High School GPA	High School GPA	High School GPA
Pell Grant Received	Pell Grant Received	Pell Grant Received
Enhanced ACT Composite	Enhanced ACT Composite	Enhanced ACT Composite
High School: ACT Record	High School: ACT Record	High School: ACT Record
First Generation Status	First Generation Status	First Generation Status
Gender	Gender	Gender
Original Place of Residence	Original Place of Residence	Original Place of Residence
HS English Composition units	HS English Composition units	HS English Composition units
HS Lab Science units	HS Lab Science units	HS Lab Science units
HS Social Science units	HS Social Science units	HS Social Science units
HS High School State: ACT Record	HS High School State: ACT Record	HS High School State: ACT Record
Ethnicity	Ethnicity	Ethnicity
HS GPA on "good" units	HS GPA on "good" units	HS GPA on "good" units
HS Lack of Foreign Language	HS Lack of Foreign Language	HS Lack of Foreign Language
Current Residency Status	Current Residency Status	Current Residency Status
Recent High School Graduate	Recent High School Graduate	Recent High School Graduate
		ACT Reading
		HS Lack of Math
		HS Foreign Language units

Table 2.3: Feature subsets selected by the three feature selection methods.

ber of features that give the minimum CV-MSE was 23 using linear regression methods, but 37 using lasso regression. This further highlights lasso regression's underperformance. It should, however, be noted that lasso has considerably better time performance.

Table 2.3 lists the selected pre-institutional feature subsets for each feature selection method. As stated earlier, binarization of categorical features expands it into multiple binary features. The Table 2.3 lists categorical features that have at least one binarized feature in the selected subset. The forward selection and backward elimination produce the same feature subset for the minimum CV-MSE values respectively, however, they differ in the binarized columns of the selected categorical features. The lasso regression selects ACT Reading, HS Lack of Math and HS For-

eign Language units features in addition to the features selected by the other two methods. The feature subset selected by Forward selection in combination with institutional features like semester GPA, semester credit hours, etc. is used to predict the delay in student graduation. The models used are presented in Chapter 3.

Chapter 3

Classification

According to Tang et al. [52], *Classification* is the problem of identifying to which set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances), whose category membership is known. The mathematical learning models that are constructed to solve this problem are known as classifiers. Duda et al. [14] says that, because a perfect classification performance is almost impossible, a more general task is to determine the probability for each of the possible categories. The abstraction provided by the feature-vector representation of the input data enables the development of a largely domain-independent theory of classification.

In this work, we used a multi-class classification model, wherein each instance can have multiple labels associated upon classification, to predict delay in graduation. The feature set consists of institutional as well as pre-institutional student characteristics. The pre-institutional features selected by forward selection, as listed in Table 2.3, were used.

3.1 Theory

3.1.1 Deep Boltzmann Machines

Deep Boltzmann Machines are machine learning structures composed of multiple layers of nodes that store latent variables usually called hidden units. The structure is identical to the one of the multilayer perceptron (see e.g. [20]). Every layer is usually known as a Restricted Boltzmann Machine (RBM) [22], a structure with a visible layer of nodes $\mathbf{v}$ and a layer of hidden nodes $\mathbf{h}$. A node i in the visible layer is connected to a node j in the hidden layer through the weight $w_{i,j}$. A stack of different Boltzmann machines can be constructed where the hidden nodes of a layer are the visible nodes of the following one.

Each layer is assumed to be able to store a state that depends on the previous layer. If the states are binary, a representation of the probability of each node state can be modeled as a Bernoulli distribution through a sigmoidal function of the state probabilities of the previous layers. If the used sigmoid is a logistic function, for node j of layer k, the probability of its state is

$$p(h_j^{(k)}|\mathbf{v}^{(k)}) = \frac{1}{1 - e^{-\mathbf{w}_j^{(\cdot,k\ \top)}\mathbf{v}^{(k)} + b_j}} \tag{3.1}$$

where $\mathbf{v}^{(k)} = \mathbf{h}^{(k-1)} =$ are the visible nodes of layer k, corresponding to the hidden nodes of layer $k-1$ and $\mathbf{w}_{\cdot,j}^{(k)}$ is the $j-$th row of matrix $\mathbf{W}^{(k)}$ connecting the visible and hidden nodes of layer k.

Conversely, the probability of a visible node i given its corresponding hidden layer can be also modeled as

$$p(v_i^{(k)}|\mathbf{h}^{(k)}) = \frac{1}{1 - e^{-\mathbf{w}_{i,\cdot}^{(k)}\mathbf{h}^{(k)} + c_i}} \tag{3.2}$$

where $\mathbf{w}_{i,\cdot}$ is the $i-$th row of matrix $\mathbf{W}^{(k)}$

G. Hinton [23] showed that a DBM can be trained layer-wise in such a way that every layer represents a representation of the data in the previous layer with a higher level of abstraction. Specifically, for each layer, one can construct a joint probability distribution

$$p(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(-E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})\right) \tag{3.3}$$

where $Z(\boldsymbol{\theta})$ is a normalization factor and

$$E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) = -\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{v}^\top \mathbf{b} + \mathbf{h}^\top \mathbf{c}\right) \tag{3.4}$$

is the energy function [5], where elements of $\mathbf{W}$ represents the importance of that a given visible node i and a given hidden node j are simultaneously active, and the elements of vectors $\mathbf{b}$ and $\mathbf{c}$ represents the importance of the activation of each one of the visible and hidden nodes themselves.

The maximization of function (3.3) can be achieved by gradient descent with respect to the weights of the model. This training can be given using a greedy algorithm usually called Contrastive Divergence [23, 21, 6]. The training consists on a gradient descent with the form

$$\mathbf{W}[n] = \mathbf{W}[n-1] + \mu\mathbb{E}(\mathbf{v}\mathbf{h}^\top|\mathbf{x}_l) - \mathbb{E}(\mathbf{v}\mathbf{h}^\top) \tag{3.5}$$

The first expectation of the equation can be computed as follows. Using the input data $\mathbf{x}_l$, compute the hidden state $\mathbf{h}$ probability with equation (3.1). Then, set a value $h_j = 1$ with probability $p(h_j^{(k)}|\mathbf{v}^{(k)})$ for each node. Compute the average of the outer product $\mathbf{v}\mathbf{h}^\top$, where $\mathbf{v}$ is changed by all available data $\mathbf{x}_l$. This is the so called clamped phase of the training.

Then, in the unclamped phase, all synthesized values of $\mathbf{h}$ are used to generate likewise values of the visible layers using equation (3.2) and compute again the

average, but this time with synthesized visible data. In the convergence, both expectations tend to be equal. Since these expectations are the cross correlation between the visible and hidden nodes, in the convergence, both posteriors (3.1) and (3.2) are maximized at the same time. This means that the probability distributions have the same properties.

This training can be performed layer by layer, using the available data for the visible nodes of the first layer and then using the hidden nodes of each layer as inputs to visible nodes for the next one. Each layer will contain a feature extraction of the previous layer with increased level of abstraction. For example, if different clusters representing different classes are present in the input data, this distribution will be preserved at the output.

When the network is used as a classifier, a supervised backpropagation training is applied as a fine tuning.

3.1.2 Support Vector Machine

A Support Vector Machines (SVM) use a classification criterion based on the so called margin maximization. Assuming a classifier based on the sign of the linear estimator

$$f(\mathbf{x}_l) = \mathbf{w}^\top \mathbf{x}_l + b \tag{3.6}$$

the SVM idea consists of the minimization of a linear cost over the estimation error plus a term that maximizes the distance between the so called classification margins, which are the hyperplanes $\mathbf{w}^\top \mathbf{x} + b = \pm 1$. This is equivalent to minimize the norm of the parameter vector $\mathbf{w}$. The corresponding problem can be written as

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{b}} \|\mathbf{w}\|^2 + C \sum_l \xi_l \\ &\text{subject to } y_l \left(\mathbf{w}^\top \mathbf{x}_l\right) = 1 - \xi_l \end{aligned} \tag{3.7}$$

Solving the minimization problem using Lagrange optimization leads to the dual problem

$$\begin{aligned} &\frac{1}{2}\sum_{l,m}\mathbf{x}_l^\top\mathbf{x}_m + \sum\alpha_m \\ &\text{subject to } 0 \leq \alpha \leq C \end{aligned} \tag{3.8}$$

which can be solved using quadratic programming [43]. An important characteristic of these machines is that they have good generalization properties thanks to the maximum margin criteria, which is equivalent to the machine complexity minimization. Also, nonlinear versions can be constructed by simply changing the dot product in (3.8) by any positive definite function (often called Mercer's kernel) [4]. See [7] for a detailed description of the SVM.

3.1.3 Gaussian Process Classifier

The Gaussian Process (GP) classifier is rooted in the Bayes theory. The underlying idea of the GP classifier consists of the construction of a Gaussian prior distribution over the estimation function $p(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b$, and then constructing a probabilistic estimator using a sigmoid function that produces a prior

$$p(y = 1|\mathbf{x}) = \text{sigmoid}(f(\mathbf{x})) \tag{3.9}$$

The distribution of the latent variable y corresponding to a test sample, can be computed by using the posterior over the latent variables $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ and then, using the prior (3.9) a posterior can be computed of the form

$$p(y = 1|\mathbf{X}, \mathbf{y}, x) \tag{3.10}$$

where $\mathbf{X}, \mathbf{y}$ are the training samples, and $\mathbf{x}$ is the test sample. This inference is in general intractable, but approximation exist that asymptotically tend to an optimal solution. Also, kernel versions are straightforwards. See [27] for a detailed description.

3.2 Experimental Setup

3.2.1 Dataset

We used the actual student data from The University of New Mexico (UNM) for our experiment. We used incoming student cohorts for the years ranging from 2006 to 2010. Furthermore, only First-Time Full-Time (FTFT) students, i.e. full-time, first-time degree seeking undergraduate students, were included in this analysis. This cohort was particularly chosen so as to have sufficient data for training and also because this sample will give us the required classification labels for testing and validation. We chose the following student characteristics, listed in Table 3.1 as our feature set for the first set of experimentation (we will refer to this as Dataset 1 from here on).

Feature Set	
High School GPA	First and Second Semester GPA at UNM
Number of Credit Hours taken by a Student up to Second Semester at UNM	If the student went through a Major/Degree change upto Second Semester
Enhanced ACT Composite	HS Lack of Foreign Language
Gender	Pell Grant Received
Original Place of Residence	Current Residency Status
HS units of English Composition	High School: ACT Record
High School State: ACT Record	First Generation
Ethnicity	HS GPA on "good" units
HS Lab Science units	HS Social Science units
Recent High School Graduate	

Table 3.1: Dataset 1: Pre-institutional and institutional features used for the first set of experimentation.

The target values comprised of the following three classes:

1. **Class 1**: No delay in graduation, i.e. student graduated in four years

2. **Class 2**: Delay in graduation of one year

3. **Class 3**: Delay in graduation of two or more years

The number of students/records in the dataset used for experimentation was 16,174. The first set of experimentation takes the feature subset selected by the forward selection, as listed in Table 2.3. It further includes institutional features such as Semester GPA and Credit Hours taken up to the first year at UNM. The features for the first set of experimentation are listed in Table 3.1. In the next set of experimentation, features describing institutional characteristics of students were extended to include their second year at UNM. Accordingly, the following were added to the earlier feature set for our second set of experimentation (We will refer to this as Dataset 2 from here on):

- Third and fourth semester GPA at UNM
- Number of credit hours taken by a student up to fourth semester at UNM

Finally, we extend our experimentation to include the institutional characteristics of students up until the third year at UNM in our feature set. The following were added to the earlier feature set for our final set of experimentation (We will refer to this as Dataset 3 from here on):

- Fifth and sixth semester GPA at UNM
- Number of credit hours taken by a student up to sixth semester at UNM

3.2.2 Construction of Deep Boltzmann Machine

Our model consists of the following steps: building a neural network, training the neural network with DBM trainer, and finally using supervised backpropagation

trainer as a fine tuning. We used *softmax* layer in backpropagation for our multi-class classification.

Constructing the Neural Network

The neural network has been constructed with one input layer, two hidden layers and an output layer. This is illustrated in Fig. 3.1.

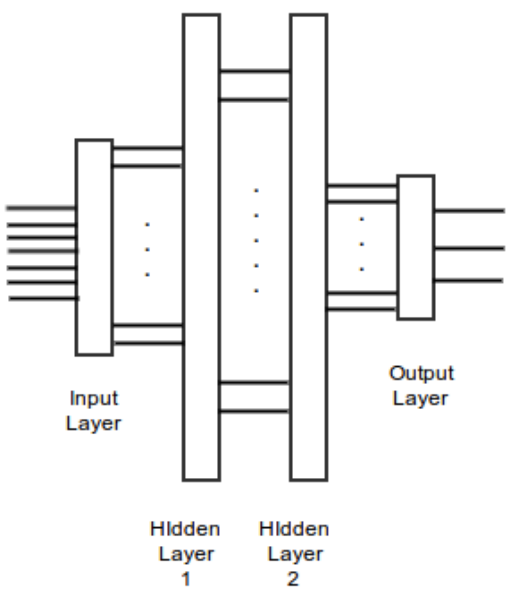

Figure 3.1: Generalized structure of the network used.

The configuration of neurons in each layer for our first set of experimentation is as follows:

Input Layer:	27
Hidden Layer 1:	256
Hidden Layer 2:	256
Output Layer:	3

The input neurons and output neurons correspond to the input features and output targets as mentioned in Section 3.2.1. Furthermore, the input layer size changes as we add features for our additional two sets of experimentation.

Unsupervised Training

The neural network constructed in the previous step can be seen as a stack of Restricted Boltzmann Machines (RBM) and are trained as explained in Section 3.1 earlier. This phase uses *reconstruction*, that is, to estimate the probability distribution of the original input, and thus is a generative process. The training parameters used in this step, that gave the best performance on cross validating, are as follows:

- learning rate = 0.05
- batch size = 32
- number of epochs for each RBM layer = 10
- activation function = Rectified Linear Units

Finetuning

The supervised backpropagation algorithm was used for fine tuning with the following parameters (these parameters were chosen as they gave the best results while cross validating them):

- learning rate = 0.1
- batch size = 32
- number of epochs = 100
- classification layer = softmax
- dropout rate = 0.1
- activation function = Rectified Linear Units

The results are presented in Section 3.3.

3.2.3 Support Vector Machines

The Support Vector Machine model was implemented for all the three sets of experiments, that is with datasets having varying degrees of institutional characteristics, as explained in the sections above. The kernel used was the Radial Basis Function. The best parameters for the SVM model were found out through cross validation. The cost parameter C was chosen as 100 and the free parameter γ was chosen as 0.1. The accuracy did not show significant variation in iterating the cost parameter C through 1 to 100 with a step size of 1, but it did vary considerably with free parameter γ, iterating through values in the range 0.01 to 1 with a step size of 0.01. The results are presented in Section 3.3.

3.2.4 Gaussian Process Classification

The Gaussian Process model for classification was implemented for all three sets of experiments, as in the previous methods. Isotropic Radial Basis Function was used

as the kernel. Herein, the classification is done based on a probabilistic approach, i.e. class probabilities obtained are translated into classification labels.

Here, the Gaussian process classifier was fitted for each combination of pairs of target classes, and is trained to do binary classification among each pair. Then, these binary predictions are combined to result into multi-class predictions.

3.3 Experimental Results

The precision, recall, and F1-score value associated with each of the target classes, namely no delay (Class 1), one year delay (Class 2), and two or more years delay (Class 3) in graduation, was calculated from the confusion matrix and averaged over the iterations of 5-fold cross validation. This is illustrated in Figure 3.2, Figure 3.3 and Figure 3.4. Figure 3.2 shows the comparison of DBM, SVM and GP based on the aforementioned measures for Dataset 1. Similar comparison has been made for Dataset 2 and Dataset 3 in Figure 3.3, and Figure 3.4, respectively. Figure 3.5 depicts the comparison of overall mean accuracy of 5-fold cross validation obtained for all the 3 datasets for each model.

As it can be seen from Figure 3.2, DBM tends to perform much better than SVM and slightly better than GP for Dataset 1, owing to the low precision, recall and F1-score values for Class 2 exhibited by SVM and GP. Furthermore, SVM highly under-performs DBM and GP for Dataset 1, as is clear from the F1-scores and error-rate values for classes 1 and 2 in Figure 3.2. Though it is interesting to note that the difference in overall accuracy for all the three models with each other is insignificant, which can be clearly inferred from Figure 3.5. Figure 3.3 shows that the precision, recall and F1-score values of individual classes increases for Dataset 2, which reflects that the addition of more institutional features improves the classification. This is also reiterated by a noticeable increase in overall accuracy illustrated in Figure 3.5.

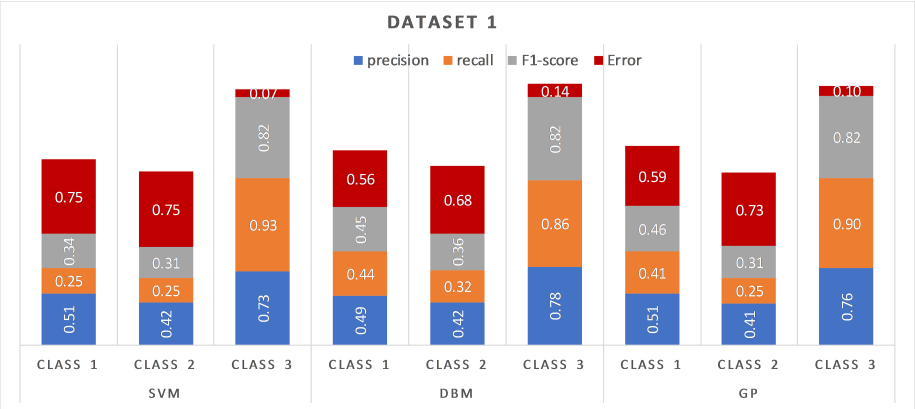

Figure 3.2: Precision, recall, F1-score, and error-rate for each class and each model for Dataset 1.

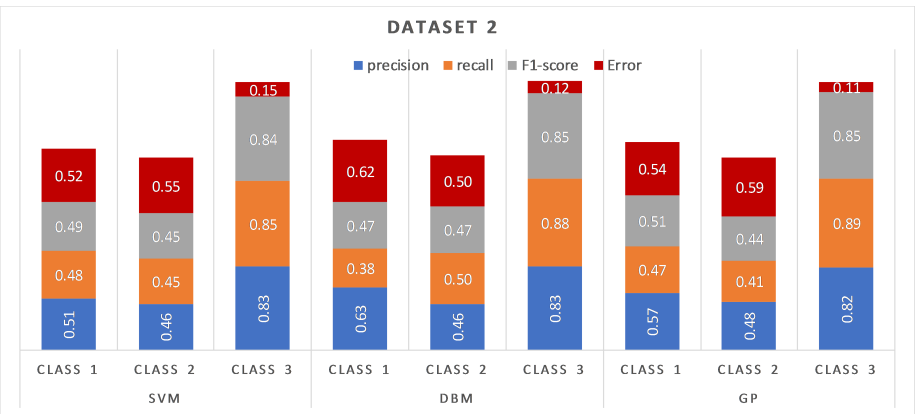

Figure 3.3: Precision, recall, F1-score, and error-rate for each class and each model for Dataset 2.

Comparing the models based on the precision, recall, F1-score, and error-rate values for each class in Figure 3.3, we find that the models have negligible differences, except for SVM, which slightly under-performs GP and DBM for classes 1 and 2. Figure 3.4 and Figure 3.5 again reflect the increase in performance for all three models with the addition of more institutional features in terms of the aforementioned metrics. According to Figure 3.4, all three models can be said to show almost negligible differences in performance.

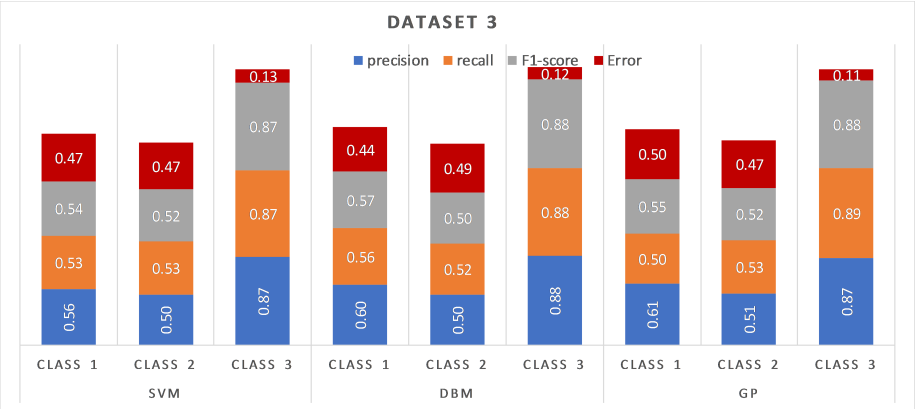

Figure 3.4: Precision, recall, F1-score, and error-rate for each class and each model for Dataset 3.

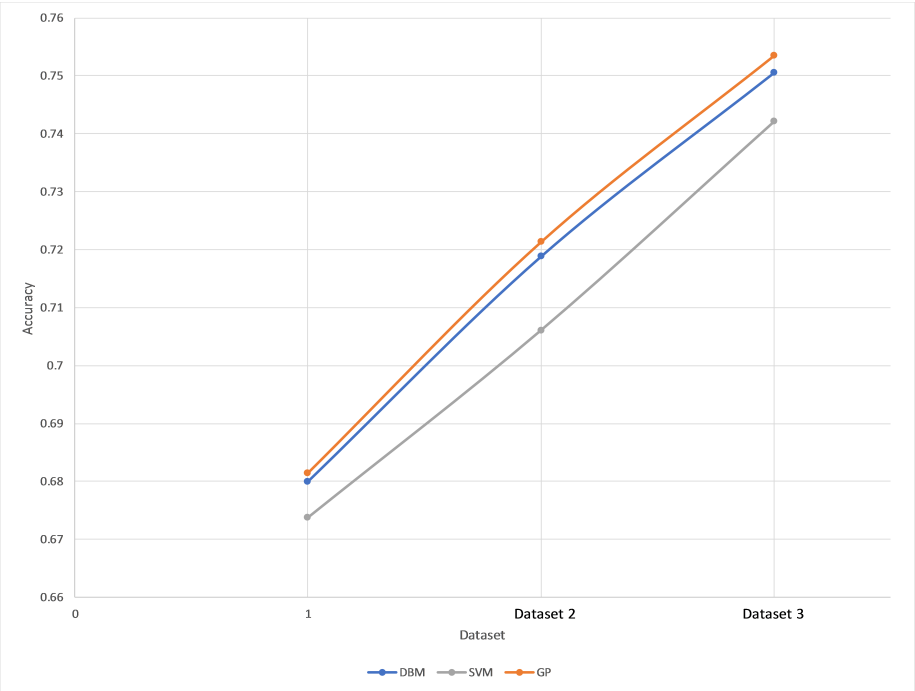

Figure 3.5: Overall accuracy obtained for DBM, SVM, and GP for each dataset.

The comparatively low precision, recall, and F1-score values for class 2 with respect to class 1 and class 3 for all the 3 datasets can be attributed to some extent to its low number of records/students in the dataset. Comparing the models based

on the time for computation, given the same computational resources and dataset, SVM performs the best and GP performs the worst. Thus, overall DBM can be said to be more suitable in terms of better classification of each class for each dataset as compared to SVM, and better in terms of computation time with respect to GP.

3.3.1 Balancing the Classes

As stated above, low precision and recall values for classes 1 and 2 can be because of their smaller sample size as compared to class 3. A possible way to mitigate this problem is to take the individual class weights into account.

We weighted SVM's cost parameter, **C**, by multiplying it with inverse class frequencies, as follows:

1. for class 1: $w_1 = n/n_1$
2. for class 2: $w_2 = n/n_2$
3. for class 3: $w_3 = n/n_3$

where, n_1, n_2, and n_3 are the class frequencies for classes 1, 2, and 3, respectively.

Figure 3.6 compares the error rate and F1-score for each class for class weighted (CW) SVM and non class-weighted SVM. It can be clearly seen that the error rate decreases and F-1 score increases for classes 1 and 2 while using class-weighted SVM. The error rate values and F1-scores for each class are averaged over 5-folds of cross validation.

Figure 3.6: Error-rate and F1-score for each class for class-weighted (CW) SVM and non class-weighted SVM using Dataset 1.

Chapter 4

Conclusions and Futurework

We presented three feature selection methods, Forward Selection with Linear Regression (wrapper approach), Backward Elimination with Linear Regression (wrapper approach), and Lasso Regression (embedded approach). These feature selection methods were used to choose the best subset, i.e., the feature subset having the lowest CV-MSE, of pre-institutional features. We used real UNM student data for our experimentation. The two wrapper approaches had similar performance in terms of CV-MSE. However, they, had a lower CV-MSE for the best performing feature subset, compared to that for lasso. High school GPA, ACT scores, student's high school, financial aid received, first generation status, and current residency status were found to be important for predicting success.

We presented three predictive models, DBM, SVM and GP, to predict the no delay, one year delay, and two or more years delay in student graduations. The feature subset selected by the forward selection method, combined with institutional features, semester GPA, semester credit hours, and change in major, up to the first two semesters, were used to predict the delay in graduation of a student for the first set of experimentation. We extended our experiments to two more sets, adding

more institutional features, such as semester GPA and cumulative credit hours to the earlier dataset. This resulted in a noticeable increase in overall accuracy as more institutional features were added. In terms of overall accuracy, the difference in performance between all three models was negligible. DBM performed slightly better overall than SVM in terms of precision, recall, F1-score, and error-rates for individual classes, whereas DBM and SVM performed better than GP in terms of computational time given the same resources. The performance of GP was comparable to DBM and better than SVM for the individual classes. Furthermore, the comparatively diminished performance of class 1 and 2, with respect to class 3, can be partially attributed to their lower percentage in the dataset. This was addressed for the case of SVM by taking into account the class weights. This resulted in a significant decrease in error-rates and increase in F1-scores for classes 1 and 2.

The models can predict delay in graduation with overall accuracies in the region of 68%, 72% and 75% for Dataset 1, 2 and 3 respectively. This suggests that there exist other pre-institutional and institutional factors that should be included in the feature set. Furthermore, class weights need to be taken into account for DBMs and GPs to improve the performance of the underrepresented classes, 1 & 2. These will be the focus of our future work.

Bibliography

[1] Education knowledge and skills for the jobs of the future. The White House. Available at `https://www.whitehouse.gov/issues/education/higher-education`.

[2] Access and equity report. *Chronicle of Higher Education (CHE)*, 58(1):42–49, 2011-12.

[3] Predictive analytics for student success: Developing data-driven predictive models of student success. *Technical report, University of Maryland University College*, January 2015.

[4] N Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(3):337–404, May 1950.

[5] Marc'aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann L. Cun. Efficient learning of sparse representations with an energy-based model. In P. B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1137–1144. MIT Press, 2007.

[6] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In P. B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 153–160. MIT Press, 2007.

[7] C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):1–32, 1998.

[8] Anthony P. Carnevale. The economic value of college majors executive summary 2015. Georgetown University Center on Education and the Workforce, McCourt School of Public Policy, 2015.

[9] Anthony P. Carnevale, Nicole Smith, and Jeff Strohl. Recovery: Job growth and education requirements through 2020. Georgetown University Center on Education and the Workforce, McCourt School of Public Policy, June 2013.

[10] Matthew M. Chingos and Sandy Baum. The federal-state higher education partnership. `http://www.urban.org/sites/default/files/publication/90306/2017.4.26_how_states_manage_their_roles_finalized_0.pdf`, March 2017.

[11] Linda DeAngelo, Ray Franke, Sylvia Hurtado, John H. Pryor, and Serge Tran. Completing college: Assessing graduation rates at four year college. *Technical report, Higher Education Research Institute, UCLA, Los Angeles, CA*, November 2011.

[12] Justin Doak. Cse-92-18 - an evaluation of feature selection methods and their application to computer security. Technical report, UC Davis: College of Engineering, Davis, California, March 1992.

[13] Fact sheet: Focusing higher education on student success. `https://www.ed.gov/news/press-releases/fact-sheet-focusing-higher-education-student-success`. Accessed: 2017-05-12.

[14] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[15] Jennifer G. Dy and Carla E. Brodley. Feature selection for unsupervised learning. *J. Mach. Learn. Res.*, 5:845–889, December 2004.

[16] Tiffany Dovey Fishman, Allan Ludgate, and Jen Tutak. Success by design: Improving outcomes in american higher education. Deloitte University Press. `https://dupress.deloitte.com/dup-us-en/industry/public-sector/improving-student-success-in-higher-education.html`, March 2017.

[17] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

[18] Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)*, 2003.

[19] Mark A. Hall and Lloyd A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, pages 235–239. AAAI Press, 1999.

[20] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[21] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[22] G. E. Hinton and T. J. Sejnowski. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning and Relearning in Boltzmann Machines, pages 282–317. MIT Press, Cambridge, MA, USA, 1986.

[23] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[24] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273 – 324, 1997. Relevance.

[25] George D. Kuh, Jillian Kinzie, John H. Schuh, and Elizabeth J. Whitt. *Student Success in College: Creating Conditions That Matter*. Jossey-Bass, San Francisco, CA, 2010.

[26] Vipin Kumar and Sonjharia Minz. Feature selection: A literature review. *Smart Computing Review*, 4:211–229, June 2014.

[27] M. Kuss and C. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Machine learning research*, 6:1679–1704, Oct 2005.

[28] Huan Liu and Hiroshi Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[29] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502, April 2005.

[30] Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Brief Bioinform*, 9(5):392–403, Sep 2008. bbn027[PII].

[31] Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):301–312, March 2002.

[32] Graduating on time saves you money. `http://www.utsa.edu/moneymatters/cost/graduating.html`. Accessed: 2017-05-10.

[33] From transformation-based dimensionality reduction to feature selection. In *Proceedings of the 27th International Conference on Machine Learning*, pages 751–758, 2010.

[34] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, C-26(9):917–922, Sept 1977.

[35] Fast facts: Graduation rates. `https://nces.ed.gov/fastfacts/display.asp?id=40`. Accessed: 2016-12-14.

[36] Digest for education statistics. `https://www.ed.gov/news/press-releases/fact-sheet-focusing-higher-education-student-success`. Accessed: 2016-12-14.

[37] Tuition costs of colleges and universities. National Center for Education Statistics, Digest for Education Statistics.

[38] Performance-based funding for higher education. In *National Conference of State Legislatures*, January 2015.

[39] Tushar Ojha, Gregory L. Heileman, Manel Martinez-Ramon, and Ahmad Slim. Prediction of graduation delay based on student performance. In *Proceedings of the International Joint Conference on Neural Networks*, Anchorage, Alaska, USA, 2017. IEEE.

[40] Judea Pearl. *Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley Publishing Company, Reading, MA, 1984.

[41] Indicators of higher education equity in the united states. The Pell Institute for the Study of Opportunity in Higher Education, Penn Ahead-Alliance for Higher Education and Democracy. `http://www.pellinstitute.org/downloads/publications-Indicators_of_Higher_Education_Equity_in_the_US_45_Year_Trend_Report.pdf`, March 2015.

[42] The changing profile of student borrowers. Pew Research Centers, October 2014.

[43] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 185–208, Cambridge, MA, 1999. MIT Press.

[44] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, March 1986.

[45] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[46] Camille L. Ryan and Kurt Bauman. Educational attainment in the united states: 2015. Technical report, United States Census Bureau, Suitland, Maryland, March 2016.

[47] Ahmad Slim, Gregory L. Heileman, Jarred Kozlick, and Chaouki T. Abdallah. Predicting student success based on prior performance. In *Proceedings of the 5th IEEE Symposium on Computational Intelligence and Data Mining*, Orlando, FL, 2014. IEEE.

[48] Ahmad Slim, Jarred Kozlick, Gregory L. Heileman, and Chaouki T. Abdallah. The complexity of university curricula according to course cruciality. In *Proceedings of the 8th International Conference on Complex, Intelligent, and Software Intensive Systems*, Birmingham City University, Birmingham, UK, 2014. IEEE.

[49] Ahmad Slim, Jarred Kozlick, Gregory L. Heileman, Jeff Wigdahl, and Chaouki T. Abdallah. Network analysis of university courses. In *Proceedings of the 6th Annual Workshop on Simplifying Complex Networks for Practitioners*, Seoul, Korea, 2014. ACM.

[50] Ahmad H. Slim, Gregory L. Heileman, Jarred Kozlick, and Chaouki T. Abdallah. Employing markov networks on curriculum graphs to predict student performance. In *Proceedings of the 13th International Conference on Machine Learning and Applications*, Detroit, MI, 2014. IEEE.

[51] Le Song, Alex Smola, Arthur Gretton, Karsten M. Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 823–830, New York, NY, USA, 2007. ACM.

[52] Jiliang Tang, Salem Alelyani, and Huan Liu. chapter Feature Selection for Classification: A Review, pages 37–64. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, Jul 2014. 0.

[53] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[54] Student debt and the class of 2014. The Institute for College Access and Success, October 2015.

[55] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 1(45):89–125, 1975.

[56] Andrea Venezia, Patrick M. Callan, Joni E. Finney, Michael W. Kirst, and Michael D. Usdan. The governance divide: A report on a four-state study on

improving college readiness and success. Technical report, The National Center for Public Policy and Higher Education, San Jose, CA, Sep. 2005.

[57] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461, March 2003.

[58] Sherry A Woosley and Dustin K Shepler. Understanding the early integration experiences of firstgeneration college students. *College Student Journal*, 45:700–714, 2011.

[59] Z. Xu, I. King, M. R. T. Lyu, and R. Jin. Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural Networks*, 21(7):1033–1047, July 2010.

[60] Zheng Zhao and Huan Liu. *Semi-supervised Feature Selection via Spectral Analysis*, pages 641–646.