A MACHINE LEARNING APPROACH TO PREDICT FIRST-YEAR STUDENT RETENTION RATES AT UNIVERSITY OF NEVADA, LAS VEGAS

by

Aditya Rajuladevi

Bachelor Degree in Computer Engineering

Jawahar Lal Nehru Technological University, Hyderabad, India

2014

A thesis submitted in partial fulfillment of

the requirements for the

Master of Science in Computer Science

Department of Computer Science

Howard R. Hughes College of Engineering

The Graduate College

University of Nevada, Las Vegas

May 2018

The Graduate College

We recommend the thesis prepared under our supervision by

Aditya Rajuladevi

entitled

A Machine Learning Approach to Predict First-Year Student Retention Rates at University of Nevada, Las Vegas

be accepted in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

Department of Computer Science

Fatma Nasoz, Ph.D., Committee Chair

Laxmi Gewali, Ph.D., Committee Member

Justin Zhan, Ph.D., Committee Member

Magdalena Martinez, Ph.D., Graduate College Representative

Kathryn Hausbeck Korgan, Ph.D., Interim Graduate College Dean

May 2018

Abstract

First-Year student retention rates refer to the percentage of first-year students who return to the same institution for their sophomore year. The national average in the institutions at the U.S for the year 2016 is at around 77 % which indicates that most of the universities are performing poorly in terms of retaining the first-year students. First-year retention rates act as an important indicator of the student satisfaction as well as the performance of the university. Moreover, universities with low retention rates may face a decline in the admissions of talented students with a notable loss of tuition fees and contributions from alumni. Hence it became important for universities to formulate strategies to identify students at risk and take necessary measures to retain them. Many universities have tried to develop successful intervention programs to help students increase their performance. However, identifying and prioritizing students who need early interventions still remains to be very challenging.

The retention rate at the University of Nevada, Las Vegas (UNLV) is close to 74% which indicate the need for specific intervention programme's to retain the students at risk of dropping out after their first year. In this thesis, we propose the use of predictive modeling methods to identify students who are at risk of dropping out after their first year at an early stage to whom the instructors can offer help. We apply several important classification algorithms from machine learning such as Logistic Regression, Decision trees, and Random forest classifier and evaluate them using metrics useful to the educators at UNLV. We use feature selection methods to identify important variables used in each model to increase the generalization and accuracy of the predictions. We propose to design effective metrics to evaluate the model's performance to help match at-risk students with appropriate supports. The main focus of this research is on students at risk of not being retained at the end of the first year, but it also lays a foundation for future work on other adverse academic outcomes.

Acknowledgements

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

The first-year or freshmen retention rate refers to the number of freshmen in a college or university who return for their sophomore year. Many universities are facing huge problems with low or decreasing first-year student retention rates. Low retention rates are a bad indicator of the university's performance and can damage the reputation of the institution in the eyes of students and parents. The reasons behind student dropout after the first year in universities can range from high expectations of the college programs, transition into an interdisciplinary curriculum, economic problems, inability to mix well with other students or struggling due to unfulfilled prerequisite requirements [Lau03]. Many researchers have formulated solutions such as building learning communities, providing additional resources [Tin99], highlighting student participation in campus life and providing academic support [Lau03]. Also, few studies have indicated that the risk of dropping out decreases with an increase in academic performance [AMBS99]. Thus, one way to increase retention is to increase academic success. In recent years, many universities have invested significantly in development and implementation of intervention programs to help at-risk students and support them individually to improve their academic performance.

The success of such intervention programs depends on the university's ability to accurately identify students who need help. In a traditional approach, many universities have used academic performance indicators such as GPA's, absence rates, previous grades, SAT or ACT scores from enrollment data to generate rules that can be used to identify students at risk [BS16]. Although such rule-based systems served as good indicators of identifying at-risk students for some years, they had some downsides such as fewer accuracies, static, expensive to generate and maintain and most importantly they lacked a validation mechanism to verify the predictions. Alternatively, recent research has indicated the potential value of machine learning algorithms such as Logistic

1
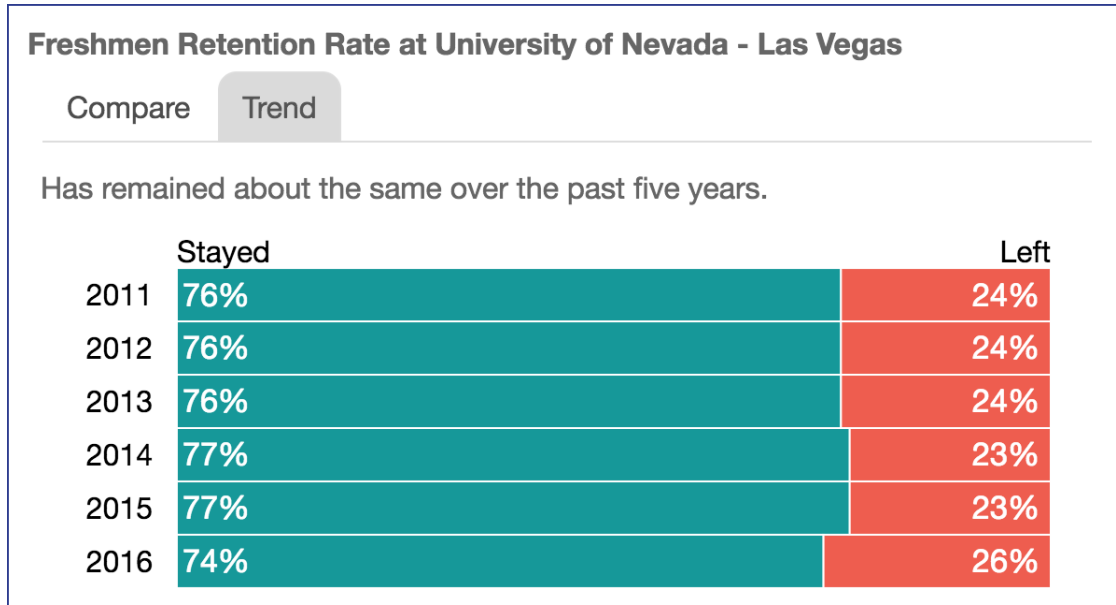
Figure 1.1: Freshmen Retention Rates at UNLV

Regression, Random Forest Classifiers, Decision Trees, Support Vector Machines (SVM) and Neural networks for the problem [Pla13, LAS+15, MDDM16]. These algorithms when trained using traditional academic data can identify at-risk students more accurately. The performance of these algorithms can be evaluated using various metrics such as Precision, Recall, and Area Under Curve (AUC) thus giving us a good indicator to validate the results. However, the application of such predictive methods to identify at-risk students is still at its early stages, owing to the implementation complexity and the availability of data. Currently, many universities have defined rules in the collection of data to use for such a research.

Over the recent years, the retention rates at UNLV have displayed a highly varying pattern. It dropped from 77% in 2012 to 74% in 2014 and then increased to 77% in 2015, which later on fell to 74% in 2016 Figure 1.1. Such an unstable pattern of freshmen retention rates has drawn a lot of attention by the educators and administration at UNLV. Hence, a predictive approach to identifying at-risk students and supporting them with additional resources would be quite beneficial to increase the retention rates at UNLV.

1.1 Objective

The objective of this thesis is to create predictive models that can be used to identify at-risk students. In this thesis, important machine learning algorithms such as Logistic Regression, Decision

trees, Random Tree Classifiers and SVM's will be trained on real-time student data obtained from UNLV's enrollment census. The trained models will then be used to predict at-risk students from a test dataset which the model has not seen earlier. The models are evaluated and compared using metrics such as precision, recall, and area under the curve to determine which model provides the best results. The results of the analysis such as the evaluation metrics will be converted to risk scores which can be easily understood by the educators and administrators at UNLV. Another contribution of this thesis is to rank students based on risk scores which will be very helpful in an efficient allocation of resources as part of the intervention programs.

1.2 Outline

In Chapter 1, the brief topic of First-Year student retention rates, its importance to universities and the proposed approach to increase the retention rates at UNLV was described.

In Chapter 2, we will discuss the existing research on improving first-year retention rates and the background information required to understand the proposed predictive approach using machine learning. It will also cover the most important and popular algorithms of machine learning.

In Chapter 3, we will describe the methodology adapted for the analysis.

In Chapter 4, we will present the experimental results. The characteristics of datasets used for these methods will also be described.

In Chapter 5, we will summarize the proposed methods and their results along with the possible extension of this thesis.

Chapter 2

Background and Preliminaries

2.1 Related Work

The prediction of first-year student retention rates and identification of students at risk of not being retained has been a well-researched problem in the area of higher education sector for decades. Early studies involved learning the important factors that lead to student dropout by developing a theoretical model. Tinto is one of the major and earliest researchers in this area. Tinto's student engagement model [Tin99] has served as the basis for a large number of theoretical studies [Bra02]. Similar research was carried out by Ernest Pascarella, Patrick Terenzini, and Alexander Astin, which focused more on the external factors such as the institution's administration and its policies when determining the reasons for student retention [A$^+$12]. Tinto in his 2006 study [Tin06] has stated that there has been a huge increase in the number of businesses and organizations to analyze and help institutions with the student retention problem. Later, in the same study, he revealed that there was only little change in the retention rates even with some huge businesses helping the universities. He also described the importance of external factors such as student-faculty relationships, extracurricular program, and orientation programs for first years. Moreover, he incorporated the role of academic factors into his model to make it more suitable to the college structure [Tin06]. Astin in his Input-Environment model [A$^+$12], suggests that researchers should consider pre-college factors such as gender, race/ethnicity, family background, high school GPA as important for student retention.

In addition to understanding the factors responsible for student dropout, the researchers were interested in identifying students at risk of not being retained in order to intervene and prevent them from dropping out. Early research included usage of statistical and analytical meth-

ods such as logistic regression and discriminant analysis for predicting student retention rates [LAS+15, MDDM16, AC17]. The results from these models showed that the learning algorithms were in fact better than many existing rule based models in learning patterns from the existing student data. Educational data mining has emerged into an important field of research in studying student retention, because of its high accuracy and robustness in working with missing data [AH14]. In another study, Jay Bainbridge, James Melitski, Anne Zahradnik, Eitel J. M. Laura, Sandeep Jayaprakash and Josh Baron used fall 2010 undergraduate students data from four different sources and applied classifiers such as logistic regression, support vector machines and c4.5 decision trees for prediction and comparison purposes[BMZ+15]. The results showed that logistic regression and SVM trees provided higher classification accuracies compared to the decision trees to predict students at risk.

Serge Herzog a researcher from University of Nevada, Reno(UNR) campus has done some extensive research on student retention and graduate prediction. He used Decision trees and Neural networks to predict student retention of data from UNR[Her06]. FarshidMarbouti, Heidi A.Diefes-Dux, KrishnaMadhavan [MDDM16] have compared seven different prediction models for identifying at-risk students using in-semester performance factors (i.e., grades) and based on standards-based grading.Another similar research focused on the problem of imbalanced output class distribution in the field of student retention in which the researchers tested three balancing techniques such as over-sampling, under-sampling and synthetic minority over-sampling (SMOTE) along with machine learning algorithms.

Although predictive analysis using machine learning models have proved to be very effective in identifying at-risk students, they were still not efficient and useful to educators who wanted to develop academic support programs and resources to support the identified students. The primary reason being the lack of understanding of the metrics from the predictive models by the educators. A few researchers analyzed this issue at the high school level and came up with a framework to convert model accuracies to risk scores that could be used by the educators in the efficient allocation of the resources [LAS+15]. Though the above-mentioned framework was giving good results, it was restricted to school level data and thus cannot yield accurate results in the university level as both are very different environments with different set of factors. Hence, in this thesis we try to include such an analysis into college level studies of UNLV and compare our results to the existing approaches.

2.2 Preliminaries

2.2.1 Machine Learning Concepts

Tom Mitchell defined Machine Learning as [Mit97]:

> "A computer is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Example: playing chess.

E = the experience of playing many games of chess with different people

T = the task of playing chess.

P = the probability that the program will win the next game.

In general machine learning tries to learn patterns inherent in the underlying data and remembers it as experience, which it uses for predictions. It was conceptualized from the notion of learning process adopted by a human brain. Just as the human brain gets better at a specific task by repeated learning and previous experiences, a computer will also learn more patterns hidden in the data based on its previous experiences. Additionally the huge processing power of computers enables them to perform such a learning process in identifying patterns from complex data which can be very difficult for a human to understand.

2.2.2 Predictive Analytics

Predictive analytics primarily deals with extracting patterns using machine learning models from existing datasets. the extracted patterns are used to predict future events and behaviors in previously unseen data. Predictive analytics is being used in wide range of fields such as education, finance, automobile, and healthcare. The analytics is performed by running different machine learning algorithms on previously collected data. The algorithms try to learn patterns between different properties of the data and preserve the knowledge in model parameters. The resulting model is able to predict the unknown property of a future unseen data.

An illustrative example is shown in Table 2.1 which shows student data captured during admissions at a university. The aim is to predict if the student will be accepted or not by looking at the

Table 2.1: Admissions data at a University

Age	Gender	HSGPA	ACT	Accepted
20	Male	3.89	25.6	1
21	Female	3.20	22.6	1
20	Female	2.50	18.3	0
22	Male	3.10	19.2	0
19	Female	3.60	23.5	1
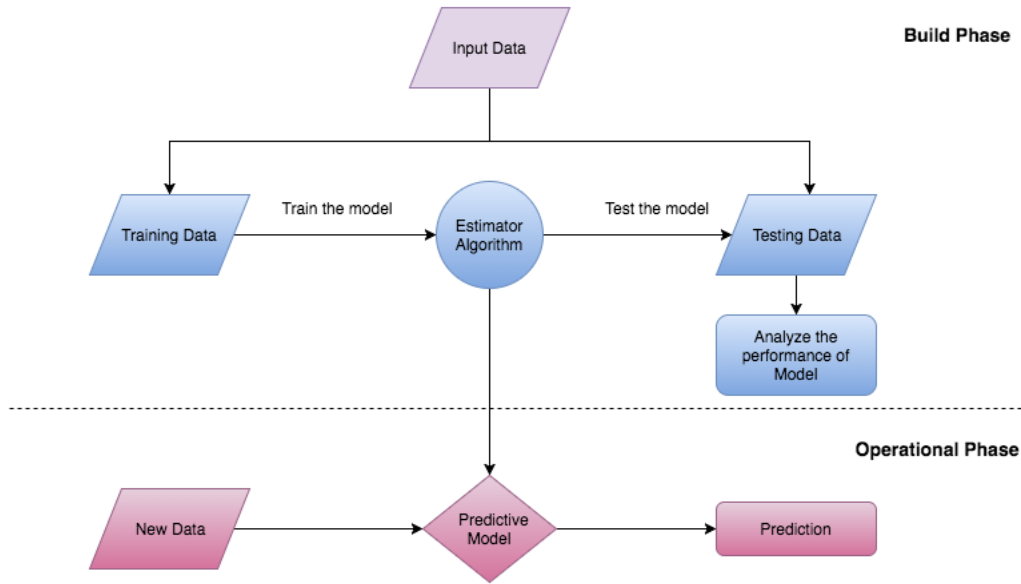
Figure 2.1: Machine Learning Approach

other variables in the data. The column "Accepted" is said to be a dependent variable, and all the other variables are called independent variables. A "1" in the "Accepted" column represents the outcome of the student being accepted into the university and a "0" means he was rejected. To this data we apply a machine learning algorithm, that will learn a prediction model based on the relations between the input independent variables and the output dependent variable. The learned model is then able to predict the output variable by taking other variables as input.

The general approach for applying predictive analysis on a data can be divided into 2 phases as shown in figure 2.1. Build phase deals with the process of creating a prediction model and testing its performance. The process of creating the prediction model is known as training and the data used is known as training data. To test the effectiveness of the created model, it is tested on another set of previously unseen data known as testing data. We use two different data sets to

check the generalization capacity of the model in predicting previously unseen data. If we use all the available data for training, the model may learn too much from the inputs giving rise to the problem of overfitting. Overfitting occurs when a model is performing well on its training data, but not on the other unseen data. A common approach to handle the problem of overfitting is to divide the input data into training and testing sets. The model is then evaluated for performance using test data. The evaluated model is used in the operational phase to perform predictions given a new data. The metrics used to evaluate a model are explained in detail in coming chapters.

2.2.3 Selected Models

In this section we discuss various models that were used in this thesis to analyze the student data from UNLV. The models selected for comparison were Logistic Regression, Decision Trees, Random Forest, and SVM's.

2.2.3.1 Logistic Regression Model

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

2.2.3.2 Decision Trees

2.2.3.3 Random Forest Classifiers

2.2.3.4 Support Vector Machines

2.2.4 Evaluation Methods

Chapter 3

Methodology

3.1 Data Collection

The Operations Support and Reporting Office at UNLV is responsible for extracting information regarding student enrollment and performance from the UNLV Data Warehouse and reporting the census to Integrated Postsecondary Education Data System (IPEDS). The data used in this thesis were obtained from the census data submitted to IPEDS every year. The following subsections will describe more information about the data.

3.2 Data Preparation

3.2.1 Data Description

The census data from the Operations Support and Reporting Office was exported into a .csv file to be used by the analysis. The raw data consisted of a total of 17,709 freshmen student records captured cohort-wise from the academic year 2012 to 2016. The data consisted of a variety of information about each student, such as pre-college academics, standardized test scores, the social and economic status of the student, freshmen year academics and finally if the student was retained in the sophomore year. Additionally, the data had variables which were not related to the first year student retention rates. After some research and insights from experts of the domain, the raw data was cleaned to only include data that would be associated with the analysis. The following subsections describe in detail about the dataset.

3.2.2 Feature Extraction

Feature extraction involves creating important useful features from the available raw data. Based on the problem statement and the data collection techniques at UNLV. The raw data was used to generate the following important features.

3.2.2.1 Computation of SAT_ACT_Score Variable

The admissions office at UNLV requires the students to report scores of standardized exams such as ACT and SAT during their admission process. From the raw dataset, we noticed that the ACT and SAT scores were missing for many students. Especially 56% of students did not have ACT scores while only 28% did not have SAT scores. Hence to capture the importance of the standardized scores into our analysis, we converted the available ACT composite scores to equivalent SAT composite scores based on the conversion table from [Hal]. The term SAT_ACT_Score was used to refer to this converted scores variable.

3.2.2.2 Computation of F2_Not_Retained Variable

The main focus of this thesis is to identify the students who are at the risk of not being retained after their freshmen year. In the raw dataset the flag F2_Retain captures the aspect of student being retained and hence can serve as the outcome variable. We then compute the complement of this flag to obtain the feature F2_Not_Retained which takes 0 if the student was retained and 1 if the student was not retained. The problem of identifying students who are at the risk of not being retained after first year is then a binary classification task with F2_Not_Retained as the outcome variable.

3.3 Data Preprocessing

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Therefore, certain steps are executed to convert the data into a tiny clean dataset. This technique is performed before the execution of any predictive Analysis.
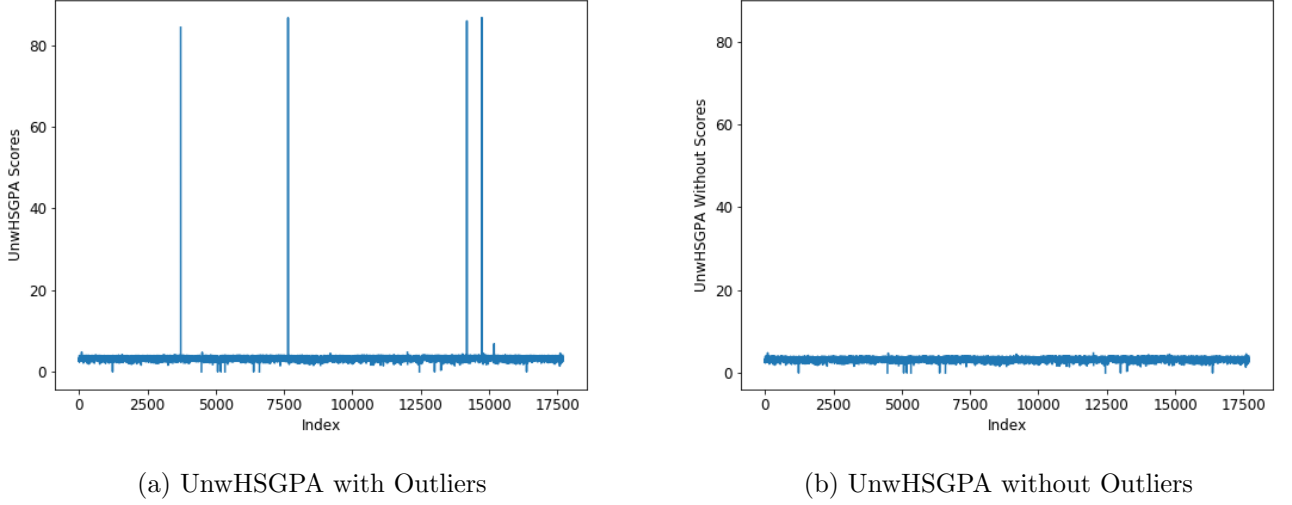
(a) UnwHSGPA with Outliers   (b) UnwHSGPA without Outliers

Figure 3.1: Removing Outliers from UnwHSGPA

3.3.1 Handling Outliers in Data

Generally the process of collecting data from different sources can introduce outliers into the dataset. This may be due to a faulty data extraction program or a human error. In predictive data analysis outliers can have significant effect on the predictive power of the models as they can introduce huge bias into the model. Hence it is almost always essential to check for outliers in the dataset before performing any analysis. Outliers were identified in our dataset. Data plots with index on x-axis and values on y-axis was used to demonstrate the presence of outlier in a feature.

3.3.1.1 UnwHSGPA scores

There were a few outliers identified for variable UnwHSGPA, as shown in figure 3.1a. The outliers of the UnwHSGPA were replaced by mean value of the said dataset excluding all the outliers. Data plot of the variable UnwHSGPA after removing outliers is shown in figure 3.1b

3.3.2 Handling Missing Values

In this subsection we will discuss how we handled missing values in the dataset. The table 3.2 shows variables along with the count of missing values. Machine learning deals with the usage of mathematical models on the data and hence it cannot be applied to datasets that have missing values. The general approach is to fill the missing value with a suitable value. The process of filling missing values in the data is known as Imputation. Researchers frequently use the mean, mode or

Table 3.1: Variables with missing values and their count

Variable	Number of missing values
SAT_ACT_Score	830
UnwHSGPA	1933
CoreHSGPA	3463
F1_MidtermGPA	286

Table 3.2: Not Retained probabilities for SAT_ACT_Score variable with respect to the availability of data points

Variable	Number of missing values
SAT_ACT_Score	830
UnwHSGPA	1933
CoreHSGPA	3463
F1_MidtermGPA	286

median of the observed given values to substitute in the missing field.

3.3.2.1 Imputing SAT_ACT_Score

The SAT_ACT_Score variable had 830 missing values. By looking at the table, it can be inferred that presence of missing values in SAT_ACT_Score variable had very little impact on the percentage of students not being retained. A total of 24.1% students were not retained when a value was present in the variable and 25.6% were not retained when the variable had a missing value.

Moreover, the boxplot of the SAT_ACT_Score vs F2_NotRetained (Figure 3.2) also shows that SAT_ACT_Score has very low impact on the student dropout. Based on the above facts, the missing values in this variable were replaced by the mean value.

3.3.3 Data Selection and Transformations

What important data did you select for your analysis

3.3.4 Correlation of Features
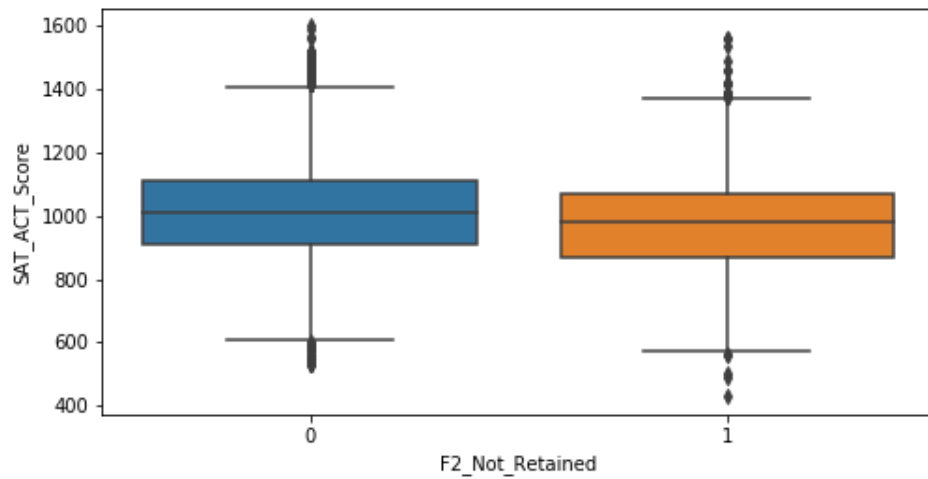
check the correlation of factors

Figure 3.2: Boxplot of the SAT_ACT_Score vs F2_NotRetained

3.4 Exploratory Data Analysis

Explain about the plots for each academic years. Explain about ipeds race and not retained Explain about mom and dad's education level and its influence on students retention

find out more about GPA's, SAT and their dependence on not retained.

Chapter 4

Experimental Results

4.1 Analysis of Predictive Models

4.1.1 Evaluation using Traditional Metrics

Chapter 5

Conclusion

Bibliography

[A+12]   Alexander W Astin et al. *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education.* Rowman & Littlefield Publishers, 2012.

[AC17]   Olugbenga Adejo and Thomas Connolly. An integrated system framework for predicting students?academic performance in higher educational institutions. *International Journal o f Computer Science and Information Technology (IJCSIT)*, 9(3):149–157, 2017.

[AH14]   Ruba Alkhasawneh and Rosalyn Hobson Hargraves. Developing a hybrid model to predict student first year retention in stem disciplines using machine learning techniques. *Journal of STEM Education: Innovations and Research*, 15(3):35, 2014.

[AMBS99]   Paul A. Murtaugh, Leslie Burns, and Jill Schuster. Predicting the retention of university students. 40:355–371, 06 1999.

[BMZ+15]   Jay Bainbridge, James Melitski, Anne Zahradnik, Eitel JM Lauría, Sandeep Jayaprakash, and Josh Baron. Using learning analytics to predict at-risk students in online graduate public affairs and administration education. *Journal of Public Affairs Education*, pages 247–262, 2015.

[Bra02]   John M Braxton. Introduction to special issue: Using theory and research to improve college student retention. *Journal of College Student Retention: Research, Theory & Practice*, 3(1):1–2, 2002.

[BS16]   Melissa A Bingham and Natalie Walleser Solverson. Using enrollment data to predict retention rate. *Journal of Student Affairs Research and Practice*, 53(1):51–64, 2016.

[Hal]   Edwards Halle. SAT / ACT Prep Online Guides and Tips.

[Her06]   Serge Herzog. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New directions for institutional research*, 2006(131):17–33, 2006.

[LAS+15]   Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, Rayid Ghani, and Kecia L Addison. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1909–1918. ACM, 2015.

[Lau03]   Linda K Lau. Institutional factors affecting student retention. *Education-Indianapolis then Chula Vista-*, 124(1):126–136, 2003.

[MDDM16] Farshid Marbouti, Heidi A Diefes-Dux, and Krishna Madhavan. Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, 103:1–15, 2016.

[Mit97]   Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[Pla13]   Mark Plagge. Using artificial neural networks to predict first-year traditional students second year retention rates. In *Proceedings of the 51st ACM Southeast Conference*, page 17. ACM, 2013.

[Tin99]   Vincent Tinto. Taking retention seriously: Rethinking the first year of college. *NACADA journal*, 19(2):5–9, 1999.

[Tin06]   Vincent Tinto. Research and practice of student retention: What next? *Journal of College Student Retention: Research, Theory & Practice*, 8(1):1–19, 2006.

Curriculum Vitae

Graduate College

University of Nevada, Las Vegas

Aditya Rajuladevi

Degrees:

Bachelor Degree in Computer Engineering 2014

Jawaharlal Nehru Technological University, Hyderabad, India

Thesis Title: A Machine Learning Approach to Predict First-Year Student Retention Rates at University of Nevada, Las Vegas

Thesis Examination Committee:

Chairperson, Dr. Fatma Nasoz, Ph.D.

Committee Member, Dr. Laxmi Gewali, Ph.D.

Committee Member, Dr. Justin Zhan, Ph.D.

Graduate Faculty Representative, Dr. Magdalena Martinez, Ph.D.