New Orleans JAZZED about Engineering Education

Paper ID #15895

A Case Study for the Application of Data and Process Mining in Intervention Program Assessment and Improvement

Ms. Elnaz Douzali, University of Illinois, Chicago

Elnaz Douzali is a senior undergraduate researcher at the University of Illinois at Chicago. She's a part of the Mechanical and Industrial Engineering Department and will receive her Bachelors of Science in Industrial Engineering in May 2016. Since 2015 Elnaz has participated in multiple projects in Educational Data Mining. Her research interests include Educational Data Mining, Process Mining, and Healthcare. Elnaz will begin her Masters of Science in Industrial Engineering at the University of Illinois at Chicago in the fall of 2016.

Dr. Houshang Darabi, University of Illinois, Chicago

Dr. Houshang Darabi is an Associate Professor of Industrial and Systems Engineering in the Department of Mechanical and Industrial Engineering (MIE) at the University of Illinois at Chicago (UIC). Dr. Darabi has been the Director of Undergraduate Studies in the Department of MIE since 2007. He has also served on the College of Engineering (COE) Educational Policy Committee since 2007. He is currently the Director of Analytics and Capacity Planning at the COE. Dr. Darabi is the recipient of multiple teaching and advising awards including the COE Excellence in Teaching Award (2008, 2014), UIC Teaching Recognitions Award (2011), and the COE Best Advisor Award (2009, 2010, 2013). Dr. Darabi has been the Technical Chair for the UIC Annual Engineering Expo for the past 5 years. The Annual Engineering Expo is a COE's flagship event where all senior students showcase their Design projects and products. More than 600 participants from public, industry and academia attend this event annually. Dr. Darabi is an ABET IDEAL Scholar and has led the MIE Department ABET team in two successful accreditations (2008 and 2014) of Mechanical Engineering and Industrial Engineering programs. Dr. Darabi has been the lead developer of several educational software systems as well as the author of multiple educational reports and papers. Some of these products/reports have already been launched/completed and are now in use. Others are in their development stages. Dr. Darabi's research group uses Big Data, process mining, data mining, Operations Research, high performance computing, and visualization techniques to achieve its research and educational goals.

A Case Study for the Application of Data and Process Mining in Intervention Program Assessment and Improvement

Abstract

The University of Illinois at Chicago offers an intervention program by admitting students to its Honors College. We call this program Honors Program (HP). HP offers additional, valuable resources that can positively affect the educational trajectory of an individual student. The current selection process for admission into HP or dismissal from HP is traditional. Administration views a students' current Cumulative Grade Point Average (CGPA) and does not look at the students' past. In this paper, we take advantage of the educational history of a student to make the admission or dismissal of each student from HP more effective. We use data and process mining techniques to study students CGPA traces and their HP participation history. We measure the graduation rates of students based on their CGPA traces and HP participation history. We show that it is possible to improve the graduation rate of students if HP admission/dismissal rules are designed based on CGPA traces and HP participation history rather than just the current CGPA of students. The model produced from our study creates a method for both students and administration to evaluate whether or not a student benefits from participating in HP. For students who are eligible and might benefit from HP, the model also determines the optimal entering semester to HP.

I. Introduction

The University of Illinois at Chicago (UIC) offers an intervention program for eligible students through its Honors College. We refer to this program as Honors Program (HP). HP offers additional, valuable resources that can positively affect the educational trajectory of an individual student. The current selection process for admission into HP is traditional. Administration views a students' current Cumulative Grade Point Average (CGPA) and does not look at the students' past. This paper proposes that the history of a student be taken into consideration for admission to HP.

This study proposes new admission rules through the use of Educational Data Mining. Educational Data Mining or EDM has been defined, on the educational data mining society website, to be "concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" [1]. Many researchers have studied the objectives and applications of EDM [2][3]. One of the applications of EDM is observing and understanding educational institutions' data to predict student retention [4]. Mohammadi et. al and Fike et. al show studies where the behaviors of students are observed and retention predictors are identified through EDM [5], [6]. Different researchers show methods for improving the outcome of student retention using EDM [7].

In addition to EDM, this paper implements process mining as its initial analysis step. Process mining is differentiated from data mining initially by its definition. As describes in the process mining book by Van Der Aalst et. al, "process mining is to use event data to extract process related information to automatically discover a process model by observing events recorded by some enterprise system" [8]. Van Der Aalst et. al also states that one of the objectives of using process mining models is insight for a modeler to view a process from different angles. In order to achieve this goal of process mining for this study, Disco software was used. Disco is a software that takes a process and produces the visualization that depicts the behavior of the input data [9]. The visualizations assist in better understanding the data and the process patterns, and therefore it helps in selecting the right analysis method.

EDM techniques and process mining create the opportunity to explore different processes and methods for UIC to establish its new rules for admission into HP. An understanding of UIC and its student population in this study is essential in this paper.

UIC is located in a metropolitan area with 15 different colleges. These colleges include entities such as the College of Engineering, the College of Liberal Arts and Science, and the College of Business. UIC offers HP for its undergraduate students. Students who participate in HP are provided resources such as private tutoring, small size classes, additional advisors and access to additional scholarships. The eligibility of a student and their participation in HP is dependent on their CGPA. The students are monitored each semester; if a HP participant's CGPA decreases below 3.4, the student is put onto academic probation then dropped from HP. Intervention programs at different universities similar to HP at UIC have been studied by several researchers [10],[11]. These studies do not use EDM in their analysis and as a result they only evaluate the overall effectiveness of the current intervention methods without providing intervention improvement policies.

In the next section, we discuss the problem description. Section III provides the solution approach, and in Section IV we show the results. Section V provides the conclusion.

II. Problem Description

This study observes the undergraduate population of approximately 17,952 students who entered UIC as a freshman between the academic years of 2004 and 2009. These students are denoted as NFTF for New First-Time Freshman. Of these 17,952 students, 1,391students participated in HP for a period of at least one semester.

HP is an intervention program that eligible students can participate in alongside their primary studies. For example, an engineering student who is eligible and applies can be selected to attend HP simultaneously as the College of Engineering. This simultaneous enrollment provides the additional resources that a student who is not a HP participant does not have access to. Some of the resources available to students who are HP participants include guidance into Research Assistants Programs, numerous merit and need-based scholarship, low student to advisor ratio, private computer labs and study lounges, and small size classes for general education courses. The intervention program aims to enhance HP participants' educational and career opportunities by providing high-achieving admitted students with additional resources and guidance.

Our focus in this study is on the HP's admission or dismissal policies for UIC NFTF who are between their second and fourth semester of their studies. Therefore, we do not intend to evaluate the policies that are used to admit NFTF to HP when they enter UIC. In addition, we do not evaluate the policies for admission or dismissal of students from HP after their fourth semester at UIC. The current admission rules for entering HP requires that a student's UIC CGPA be 3.4 or higher and a participant will be dropped from HP if they have 2 consecutive semesters with a CGPA of lower than 3.4. We argue that applying this rule to admit or to dismiss students from HP might not be the most effective strategy. We measure the impact of this rule by tracking the graduation rate of both HP and Non HP participants. We further derive a revised set of rules that can identify subcategories of students who benefit from the intervention program.

HP is an essential intervention program at UIC because it provides additional resources to students to help promote graduation while requiring excellent academic record. Our study objective was to create a model to help eligible College of Engineering students evaluate whether or not to join HP as well as the optimal time to join HP. Both administration and students would evaluate a students' enrollment each semester based on the findings of our study.

The initial step of this study was to visualize the behavior of our data through process mining. Following the initial step, conditional probabilities produced the statistics needed to perform hypothesis testing on proportions [13] to determine whether HP made a significant difference in a student's probability of graduation. For those classes of students whose probability of graduation was positively affected by HP, their education history traces were distinguished and used to construct the revised set of rules.

The concept behind process mining is the need for discovering, monitoring and improving processes by gathering information from event logs and traces [8]. A trace or event log defines the series of events that belong to an individual observation, in our study an individual student. An example of a student's trace would be their behavior in semester one onto semester two onto semester three. That individual student's unique behavior over time that is recorded in the university database system is a trace.

This paper outlines an assessment method that can be applied to a number of different educational interventions. One possible application involves the additional resources provided to students of a minority group and their educational presence and their likelihood of graduation. A similar application would be for the assessment of tutoring programs available for students who are failing or at risk of failing a course or failing from UIC.

III. Solution Approach

When the data for this study was first retrieved from our educational database, the unique identifiers were eliminated. The data fields retrieved were from student's individual information containing Term CGPA, Year entering the University as NFTF, Number of Credits, HP Participation Status, and whether the student graduated from UIC.

The first step of our analysis was to depict the behavior of students throughout their time at UIC. Disco, the software, was implemented to observe the data further. The dataset for students was structured to display three attributes of an individual student. The first attribute is the Satisfactory Status of a student: N for Unsatisfactory CGPA of below 3.4, S for Satisfactory CGPA of 3.4 or higher. The second attribute is Intervention Participation Status: 1 for participation in HP in the current semester, 0 for no participation in HP in the current semester. The third attribute was Enrollment Status: C indicates a student is continuing from the current semester to the following semester, D indicates the student has dropped from the university in the current semester, and G indicates the student has graduated from the university in the current semester. The behavior of an individual student's progression from their first semester to their last in terms of these three attributes is defined as a student's trace through the university. We enter the traces of all students to the software Disco to obtain a visualization of main trace groups in the University. The resulting process map from Disco is shown in *Figure 1*.

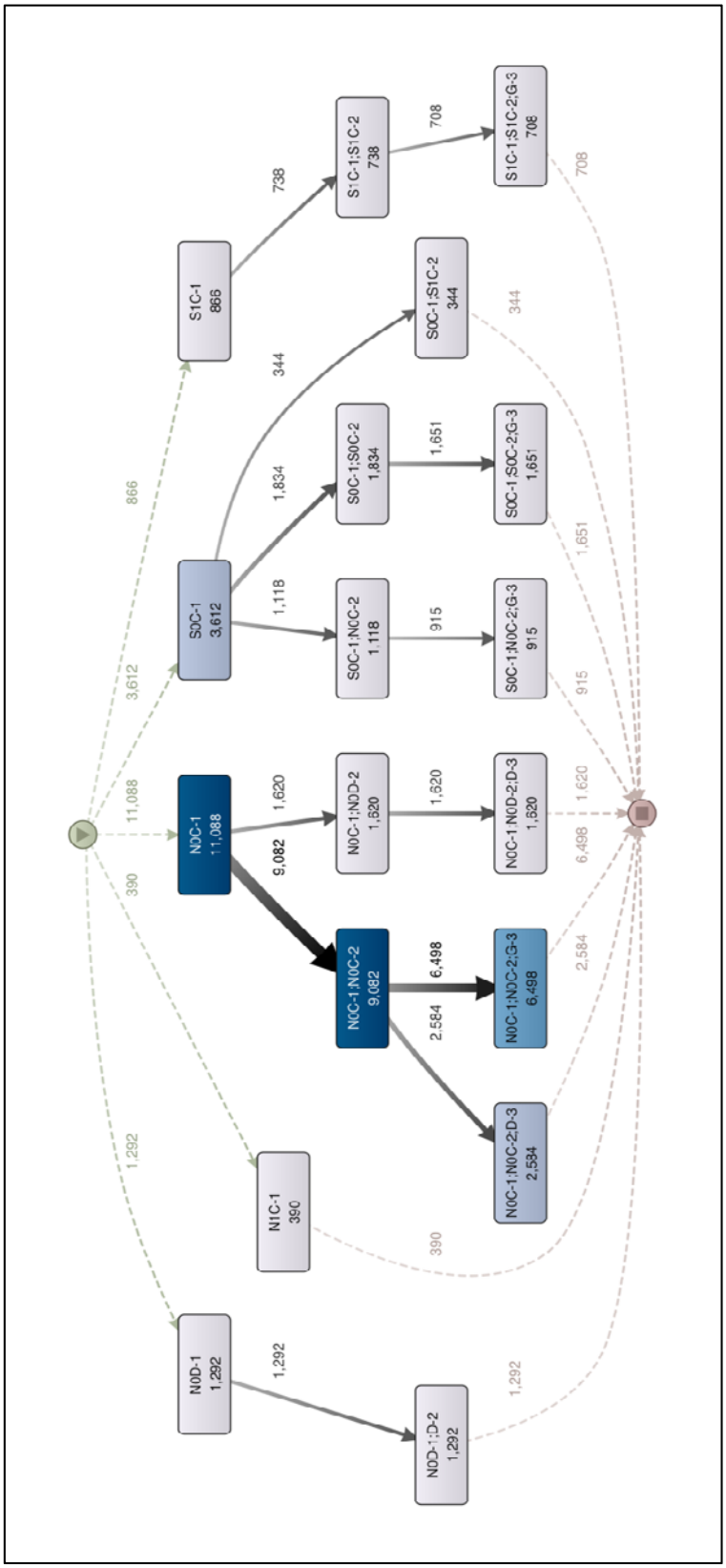
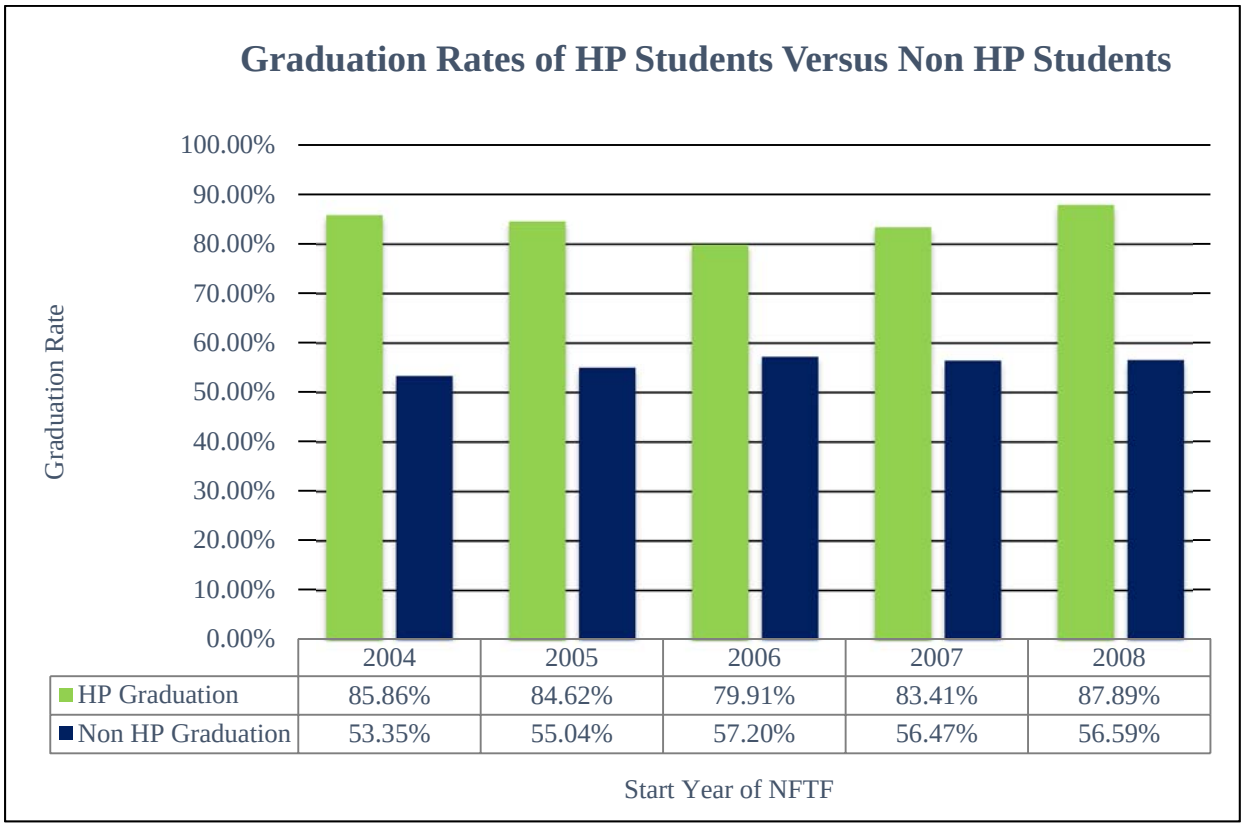


Figure 1. Disco process mining visualization of NFTF

Figure 1 illustrates how the students in our targeted population advance in the system in their first three semesters. Disco revealed that the traces of students' behavior is sequential. Process mining here identified how this problem would progress by relieving the dynamics of the problem with traces and frequencies. As an example, in *Figure* 1, it can be seen that 11,088 (node N0C - 1) students started their first semester as NFTF without participating in HP and they finished their first semester with a CGPA of below 3.4. 9,082 (node N0C – 1; N0C – 2) of these students advanced to their second semester as a Non HP participant with a CGPA of below 3.4 and they advanced to their third semester. As one can see, the summation of the outgoing flows from node N0C – 1 (9,082 + 1,620 = 10,702) is less than the input flow (11,088) to that node. This is because the Disco software was set to display the prevailing paths of student traces. All the traces that have been removed had significantly low frequencies and therefore could be considered as outlier traces.

The prevailing traces displayed in *Figure* 1 were selected for further statistical analysis. Prior to reporting the results of this statistical analysis, we would like to provide comparison statistics for graduation rates of HP participants (students who were in HP for at least one semester) versus Non HP. Non HP Participants contained 16,561 students and HP Participants contained 1,391 students. These groups were then graphed with respect to their Graduation Rates per Start Year as NFTW. The results are provided in *Figure 2*.



	2004	2005	2006	2007	2008
HP Graduation	85.86%	84.62%	79.91%	83.41%	87.89%
Non HP Graduation	53.35%	55.04%	57.20%	56.47%	56.59%

Figure 2. Graduation Rates of HP Students Versus Non HP Students

Despite a large gap in the average graduation rates between HP participants and Non HP participants, we argue that these statistics might not be a valid indicator of the performance of HP. We suggest that a better indicator of HP's performance is comparing HP participants with Non HP participants who also maintain an average CGPA of at least 3.4. These Non HP participants' CGPA may fall below a 3.4 for a semester but as long as their overall average is at least 3.4, they are included. *Figure* 3 shows the results.

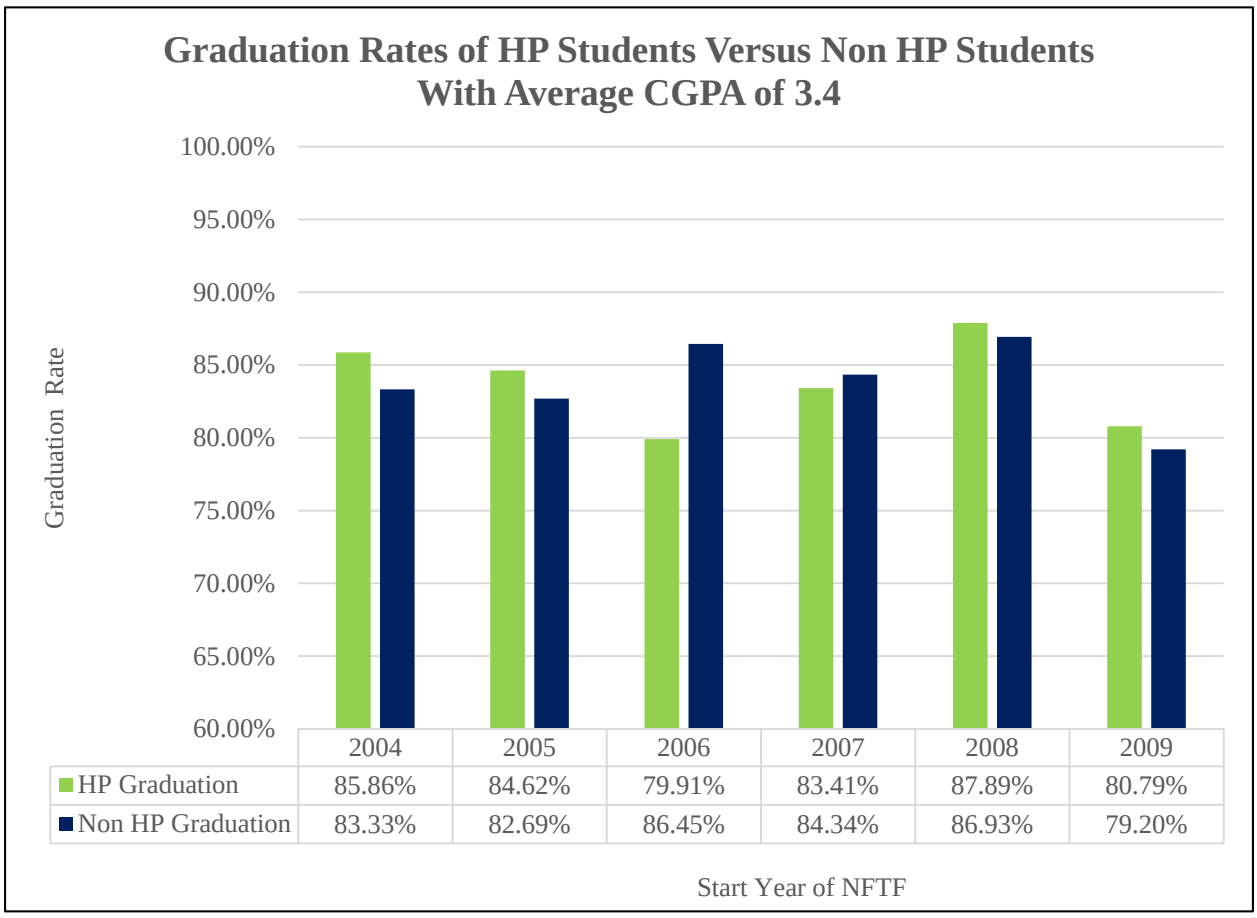


Figure 3. Graduation Rates of HP Students Versus Non HP Students with Filter

As evident from *Figure* 3, the HP participants do not necessarily perform better than Non HP participants. As evident from the figure, in 2004 HP Graduation was at 85.86% which was greater than Non HP Graduation's 83.33%. However, in 2006 the HP Graduation was 79.91% which was lower than Non HP Graduation's 86.45%. Therefore, it is necessary to perform a more detailed analysis of traces, and identify the exact classes of participants who perform better compared to their equivalent Non HP participants. In order to do this, we have provided *Figure* 4. In *Figure* 4, HP Participants include students who are in HP during each semester that they are

at UIC and therefore referred to as Strictly HP Students. The Strictly Non HP students are defined as the students at UIC who never attended HP and their CGPA was at least 3.4 every semester.

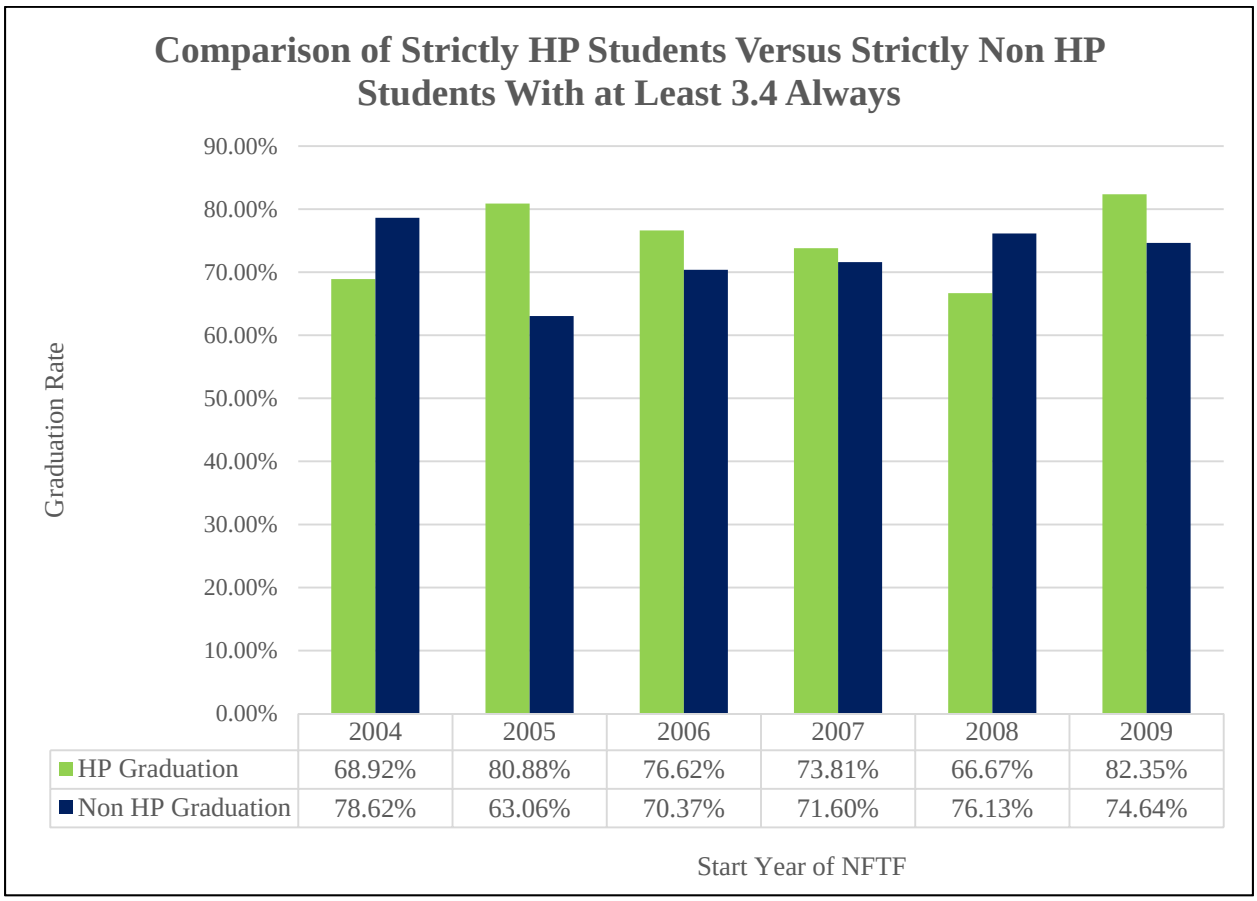


Figure 4. Strictly HP Participants and Strictly Non HP Participants of 3.4 Always

Figure 4 illustrates the behavior of graduation when comparing students who are educationally on the same level at all times. We can assume that the students who participate and the students who do not participate have equivalent opportunities of being admitted into HP. *Figure* 4 shows that students who are always in HP and students who are never in HP but always have a CGPA of 3.4 or higher perform similarly.

An additional graph is provided by *Figure* 5. This graph depicts the drop-out percentages for the combined population of HP students and Non-HP students following each semester. The graph displays both the term drop percentage and the cumulative drop percentage. *Figure* 5 verifies that the majority of students' drops are occurring in the first three semesters.

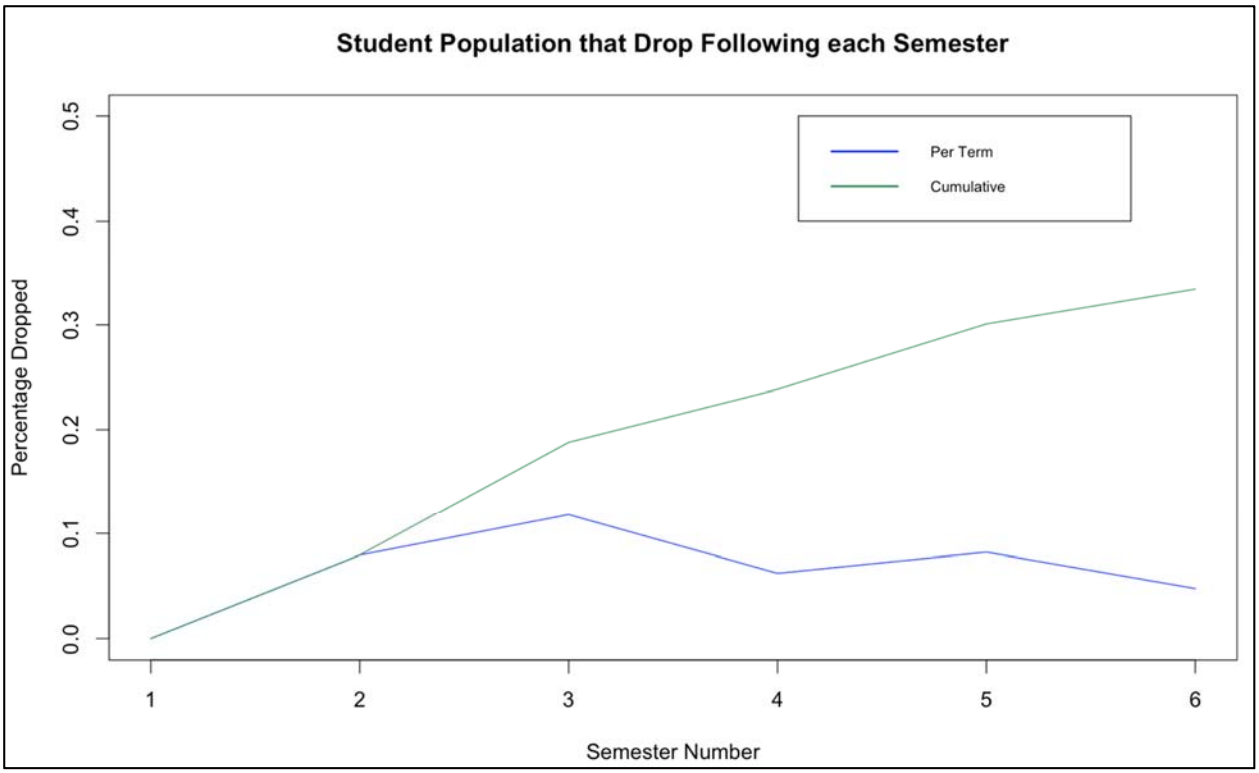


Figure 5. Combined HP and Non HP Students Drop-out Rates Per Term and Cumulative

The structure of the following analysis excludes the third attribute of a student's trace. The Enrollment Status is dropped due to the constant value of "C" in all components in the traces. The *givens* of the decision making analysis represent the component of the students' traces that have already occurred and are not subject to change. The *assigned* are whether a student of a given *trace* will participate in HP or not participate in HP, 1 and 0 respectively. Using the historical information, the frequency and probability of graduation are calculated.

Figures 6 and 7 show the details of the trace analysis and comparing the subcategories of students. Red nodes show where the selected trace subcategory does not have the appropriate sample size. Green nodes show that enough students are available to derive a decision making rule based on the information of that node. Grey nodes represent an inconclusive situation where further expansion of that node is suggested. The tables next to each node show the conditional probabilities [12] associated to that node. For example, in *Figure* 6, the node labeled by S0 can be interpreted as follows. This group of students did not participate in HP in their first semester (shown by the 0 component in S0 label) and they have a CGPA of 3.4 or higher (shown by the S component in S0 label). From the table provided next to the node S0, 492 students were assigned to HP in their second semester and 3,121 students were not assigned to HP in their second semester. The 492 students had a graduation rate of 82.52% and the 3,121 students had a graduation of 77.12%. The node S0 is labeled as green because both student sizes of 492 and 3,121 can satisfy the minimum sample size requirements for a valid statistical analysis of graduation proportions. Later in this section, we will display the formula used to determine the minimum sample size. We performed the hypothesis test of proportions for all green nodes in

Figures 6 and 7. When the nodes indicated a significant difference in the proportions (when the reported p-value in the lower part of the box next to a green node is less than 0.05), the intervention program was selected as the right decision.

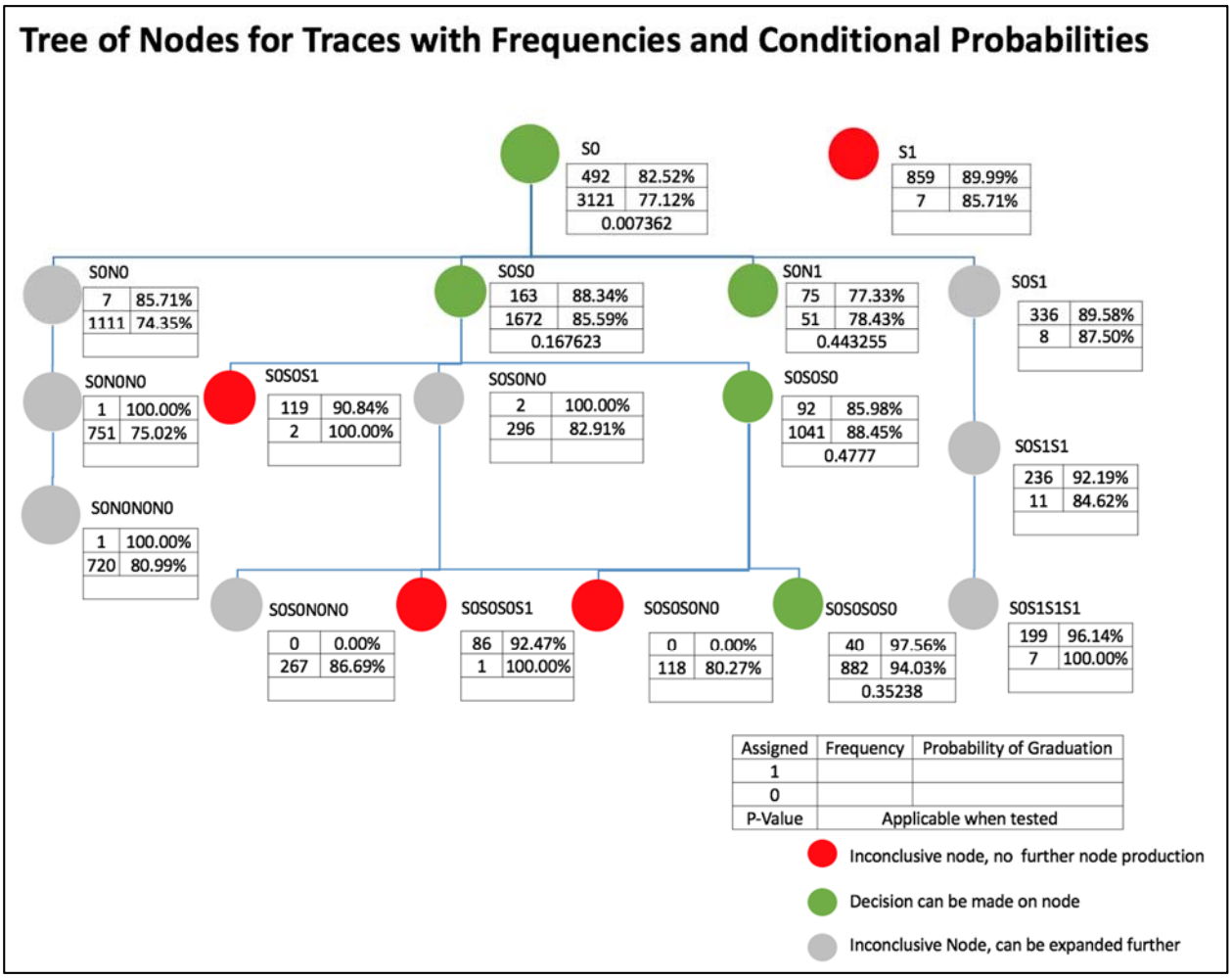


Figure 6. Semester State Tree with Frequencies and Conditional Probabilities - S0, S1

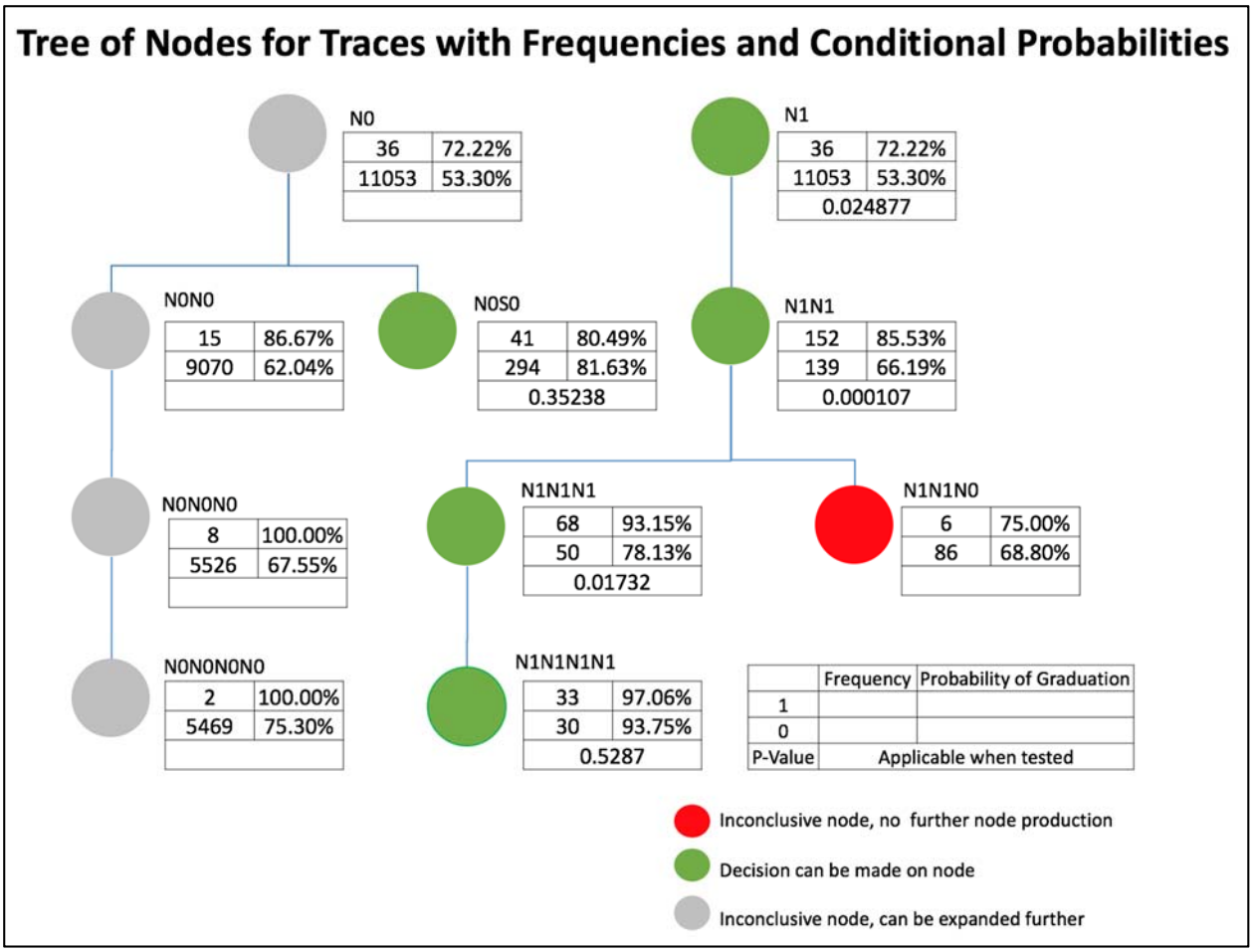


Figure 7. Semester State Tree with Frequencies and Conditional Probabilities - N0, N1

In *Figure* 7, the first node at the top is a trace of just the first component, the first semester at UIC. At this point the trace splits from N0 to N0N0 and N0S0 based on the traces that had large enough samples to expand on. For each node the frequencies and probabilities of graduation were used to determine whether or not there was a significant difference between the two populations of a node. In the case where statistical testing was appropriate based on sample size and the P-value was below 0.05, the student with that trace would be recommended to continue into HP for the following semester. For example, node N1 has a sufficient sample size to perform statistical testing. The p-value from hypothesis testing is 0.024877 which is below 0.05. A student with this trace would be encouraged to participate in HP for the following semester.

To estimate the required minimum sample sizes for each node in *Figures* 6 and 7, we used the formula given by Neibel *et. al* [13]:

$$n = \frac{Z^2_{\alpha/2} p(1-p)}{l^2}$$

where n = The number of samples needed

$Z^2_{\alpha/2}$ is Standard Normal Statistic (Z) with alpha set at 0.05

p = True percentage occurrence of population

l = Acceptable limit of error

For example, the given trace of S0S0 can be assigned to either 1 or 0. The assignment of S0S0 to 1 needs a sample size of 159 students as shown by the following equation:

$$n = \frac{3.84 \times 0.8252 \times 0.1748}{(0.05)^2} = 159$$

The assignment of S0S0 to 0 needs a sample size of 190 students as shown by the following equation:

$$n = \frac{3.84 \times 0.7712 \times 0.2288}{(0.05)^2} = 190$$

S0S0 assigned to 1 has a sample size total of 492 and S0S0 assigned to 0 has a sample size total of 3,121. This formula confirms that our dataset has the appropriate number of students to use in statistical testing. This formula confirms that the green nodes in *Figures* 6 and 7 have enough students to perform statistical testing on.

IV. Results

Based on this study's analysis, in the scenario that a student has a given history of trace N1 or S0, the individual should be assigned to the Intervention program HP for the following semester. Assigning a student to HP when it benefits their trace will increase their graduation rate based on our historic data. For example, students with a trace of N1 should be assigned to HP and that assignment is shown to increase their graduation rate of 53.30% to 72.22%. For students who are at semester 2, those who have a history of trace N1N1 should be assigned to HP. The analysis of this study produced the results for traces S1N1, S0S0, S0N1, N0N1 and N0S0. Students with these 5 traces should not be assigned to HP for the following semester. All nodes that were labeled as inconclusive or did not lead to a decision are excluded from our proposed decision making process due to the lack of sufficient data. These nodes should go about current admission procedures until further studies have occurred.

The results gathered from our study show that when students are exposed to the additional resources of HP intervention, at certain times their performance is enhanced and at certain times it is not. The objective of this study was to identify the students that would find HP program

effective. This decision making tool provides both students and administration a means of evaluating if HP will benefit a student and finding the optimal semester for a student to be subjected to HP. We have shown which traces do find HP program's intervention beneficial through the detailed analysis of our study.

V. Conclusion and Further Research

We studied the intervention program Honors Program (HP) at the University of Illinois at Chicago (UIC). Intervention program HP at UIC is an example of an offered intervention at a large university in attempts to providing resources to HP participants and promoting individual success. The objective of this study was to determine the effect of intervention program HP on the likelihood of graduation for participating students. Determining this effect allows us to improve admissions to HP, and in general making admission rules for the educational institution.

In the current intervention program, the single rule for admission into HP is that a student must maintain a CGPA of 3.4 or higher. Based on our analysis, the UIC administrators can use a set of eight different rules to admit or dismiss students from the intervention program. The creation of these rules was only possible through the application of process and data mining techniques. By the application of these rules, the administrators can allow a given subcategory of students into the intervention program only if historic evidence supports the success of that specific subcategory. This will improve the total effectiveness of the intervention.

Future work involves expanding on the variables taken into consideration for decision making. This study examined Graduation and Drop-out rates, however other factors from a student's history could improve the accuracy of the suggested rules. Additionally, a student's participation in intervention program HP in their first semester involves their behavior in high school. This study did not create rules for the first semester of a student.

References

[1] Romero, Cristobal, and Sebastian Ventura. "Educational data mining: A survey from 1995 to 2005." Expert systems with applications 33.1 (2007): 135-146.

[2] Baker, R. S. J. D. "Data mining for education." International encyclopedia of education 7 (2010): 112-118.

[3] Baker, Ryan SJD, and Kalina Yacef. "The state of educational data mining in 2009: A review and future visions." JEDM-Journal of Educational Data Mining 1.1 (2009): 3-17.

[4] Bienkowski, Marie, Mingyu Feng, and Barbara Means. "Enhancing teaching and learning through educational data mining and learning analytics: An issue brief." US Department of Education, Office of Educational Technology (2012): 1-57.

[5] Mohammadi, John. "Exploring Retention and Attrition in a Two-Year Public Community College." (1994).

[6] Fike, David S., and Renea Fike. "Predictors of first-year student retention in the community college." Community college review 36.2 (2008): 68-88.

[7] Veenstra, Cindy P. "A strategy for improving freshman college retention." The journal for quality and participation 31.4 (2009): 19.

[8] Van Der Aalst, Wil. Process mining: discovery, conformance and enhancement of business processes. Springer Science & Business Media (2011).

[9] Günther, Christian W., and Anne Rozinat. "Disco: Discover Your Processes." BPM (Demos) 940 (2012): 40-44.

[10] Blaisdell, Stephanie, and Catherine R. Cosgrove. "A theoretical basis for recruitment and retention interventions for women in engineering." age 1 (1996): 1.

[11] Belgarde, Mary Jiron, and Richard K. Lore. "The retention/intervention study of Native American undergraduates at the University of New Mexico." Journal of College Student Retention: Research, Theory & Practice 5.2 (2003): 175-203.

[12] Ross, Sheldon M. Introduction to probability and statistics for engineers and scientists. Academic Press (2014).

[13] Freivalds, Andris, and Benjamin Niebel. Niebel's Methods, Standards, & Work Design. Mcgraw-Hill higher education (2013).