Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

AN APPROACH TO POPULATION SYNTHESIS OF ENGINEERING STUDENTS FOR UNDERSTANDING DROPOUT RISK

Danika Dorris

North Carolina State University
915 Partners Way
Raleigh, NC 27695, USA

ABSTRACT

Dropping out of STEM remains a critical issue today, and it would be useful for universities to have reliable predictive models to detect students' dropout risks. Generating a synthetic population of the true population could be useful for simulating the system and testing scenarios. We outline an approach for creating a synthetic population of students in STEM using Bayesian Networks and build a microsimulation which simulates students' risk behaviors over time using Dynamic Logistic Regression prediction models. This process has identified several areas that must be addressed before the synthetic population represents the true population in a simulation.

1 INTRODUCTION

Dropping out of higher education often occurs early on in a student's academic career. If at-risk students can be identified early, then interventions could be targeted to reduce dropouts in ways that are cost-effective. Machine learning has been used to predict student dropout using a variety of university data at various points throughout a student's academic career, and population synthesis has been used in a wide range of social science applications. With a synthetic population that is representative of students over time, we are able to observe how dropout risk fluctuates over time and understand how specific university interventions may impact students. We generate the synthetic population to represent an actual student population with more than 100 different attributes accumulated over time. We match students from an historical test cohort with a representative synthetic agent and simulate dropout risk over several semesters.

2 DATA

Data from 5,348 undergraduate engineering first-year students across four undergraduate engineering cohorts was used to train our Dynamic Logistic Regression models. Similar data from 1,428 first-year students in an independent historical cohort was used to test both the prediction models and the simulation. In our analysis, we focus on students who dropout within two years from their start at the university. Data is collected on census day (last day to change the course schedule) and on the last day of the semester. Our data primarily consists of nine types of information: academic performance, community engagement, course load, demographics, financial aid, housing, admissions, residency and relatives' education level. Before the start of the first semester, there are 50 student attributes known. By the end of the of the second fall semester, we have learned approximately 160 attributes about each student.

3 METHODS

We simulated student behaviors at seven different points in time. Behaviors were determined before the first fall semester and on census day and end of each of the first fall semester, first spring semester and second fall semester. Bayesian Networks were used to generate student attributes in order to address the issue of multicollinearity in a high-dimensional data set. A Bayesian Network was first built using information known before the start of the first fall semester. This network was used to generate a population of synthetic students described by a set of attributes. As more information is revealed, a new Bayesian Network is built using both prior information and the newly revealed information. Synthetic agents are then assigned values for the new attributes inferred from their known attributes and the new Bayesian Network.

We built multiple step-wise logistic regression models using these data at each of the seven points in time. For each of the seven types of prediction, we created 100 fully balanced training sets by pairing all of the dropouts with an equal-sized bootstrap sample of non-dropouts in order to address the relatively low response rate (8 percent). The first prediction model was built using information known before the start of the first fall semester. Then, we built a model using information known on census day of the first fall semester and the average predicted likelihood of dropout across the 100 regression models built before the start of the semester. Similarly, we built the subsequent models using the information known at that time in addition to the average predicted dropout probability across all models built at the previous time step.

To establish a “true” outcome for our synthetic agents to evaluate performance of the simulation, we matched synthetic agents to actual students in the test set. Matches were made based on the Heterogeneous Euclidean-Overlap Metric (HEOM) and the outcome of the matched student from the test set became the “true” outcome for the paired synthetic agent.

Profiles for each agent were built prior to running the simulation and consist of two components: student attributes and initial risk score. The Bayesian Network built before the start of the first semester generated an agent with a set of attributes. Based on these attributes, an initial risk score is then calculated using a randomly selected prediction model built from the same time step. At the next time step, a new Bayesian network is built considering the new information and randomly assigns values for the new attributes based on the values of the known attributes. Once the new attributes are determined, a new risk score is calculated from a randomly sampled prediction model. This process is repeated at each time step to update the agent’s profile as new information is revealed.

4 RESULTS AND DISCUSSION

The means for all of the attributes of the synthetic population known before the start of the first semester fall within the 95 percent confidence interval of the attribute means for the test population. In terms of matching, 98 percent of the matches were below 2, which suggests that the agent paired with an actual student differed by less than two attributes. The predictive ability of the simulation varied widely among different time points and within a given time point, which may be caused by the variability among the logistic regression models used to calculate the risk score at each point in time. Matching students only on information known before the first semester can also limit the similarity between the agent and the student as more information is revealed in later time steps. Our analysis could be enhanced by more replications and testing multiple synthetic populations.

5 CONCLUSION

The volume of sensitive data collected by universities suggest a need for a representative synthetic population of students in order to better understand dropout risk. While we explored an approach that incorporates multiple matching methods and prediction models, our process highlighted many areas that must be considered when creating a representative synthetic population.