Dropout Prediction in Home Care Training

Wenjun Zeng*
University of Minnesota
Minneapolis, Minnesota
zengx244@umn.edu

Si-Chi Chin
SEIU 775 Benefits Group
Seattle, Washington
sichichin@gmail.com

Brenda Zeimet
SEIU 775 Benefits Group
Seattle, Washington
brenda.zeimet
@myseiubenefits.org

Rui Kuang
University of Minnesota
Minneapolis, Minnesota
kuang@cs.umn.edu

Chih-Lin Chi
University of Minnesota
Minneapolis, Minnesota
cchi@umn.edu

ABSTRACT

In Washington state (WA), SEIU 775 Benefits Group provides basic home care training to new students who will deliver care and support to older adults and people with disabilities, helping them with self-care and everyday tasks. Should a student fail to complete their required training, it leads to a break in service, which can result in costly negative health outcomes (e.g. emergency rooms and hospitalization) for their clients [1].

In this paper we describe the results of utilizing machine learning predictive models to accurately identify students who exhibit a higher risk of drop out in two areas: (1) dropping out before attending first class[first class attendance]; and (2) dropping out before completing the training[training completion]. Our experimental results show that AdaBoost algorithm gives a useful result with ROC_{AUC} = 0.627±0.013 and Precision at 10 = 0.73±0.12 for first class attendance and ROC_{AUC} = 0.680±0.024 and Precision at 10 = 0.67±0.20 for training completion without relying on additional assessment data about students. In addition, we demonstrate the use case for constructing larger decision trees to help front-line training operations staff identify intervention strategies that create the most impact in preventing dropout.

1. INTRODUCTION

By 2050, the number of Americans needing long-term home care services and supports will double[2], implying increased demand for workers providing home care services (called "personal care aides" nationally and "home care aides (HCA)" in WA). This will also increase the demand of training for HCAs to provide quality care to their clients. In WA, should an individual wish to work as a home care aide, they are required to complete a 75 hour, 2 week, Basic Training (BT) course within 120 days of their hire date. In WA, an HCA can begin providing care before completing their training as long as their deadline has not passed. In the event that an HCA fails to complete BT, she or he will fall out of compliance, leading to the HCAs termination and a break in service for the clients served by the HCA [1].

Educators have frequently used assessment tools that measure cognitive skills, engagement, self-management and social support to accurately predict student successes. However, conducting assessments at scale is time consuming for both students and instructors. In the absence of a validated assessment specific to HCA profession, there is great interest in utilizing existing learning data to isolate the strongest predictors of dropout through the predictive power of machine learning algorithms. Our research questions are two-folds: 1) Can machine learning algorithms successfully predict student dropouts? 2) What are the risk factors related to early dropout from basic home care training?

Many studies[3] have been conducted to explain academic performance and to predict the success or failure across a variety of students in a wide-range of educational settings. Machine learning algorithms have been successful in predicting graduation[4], course participation[5], and other academic outcomes[6].

However current research has not fully investigated the area of using machine learning algorithms for on-the-job training, healthcare training programs, or adult education in general. In this paper, we focus on the dropout problems in home care training using machine learning methods. We were granted the latitude to be creative with our feature engineering, utilizing readily available data to meet business requirements.

2. EXPERIMENTAL SETUP

Figure 1 illustrates the four sequential time-based milestones in home care training: 1) Complete Orientation & Safety (O&S); 2) Register for a 70-hours BT course; 3) Attend the first class in this course; 4) Complete the 70-hour training. At the moment that a prospective home care aide enters the system, a 'Tracking Date' is assigned to their O&S training

*This work has been done during the author's internship at SEIU 775 Benefits Group

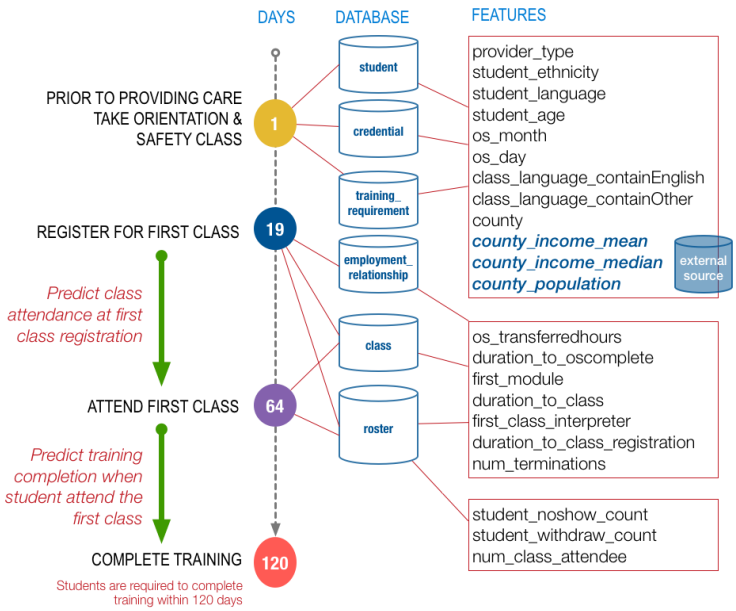


Figure 1: Predicting Targets and Features

requirement, signifying the start of their training journey. On average a student will register for his or her first class approximately 19 days after completing O&S and will actually attend his or her class about 64 days after entering our system.

Predicting dropouts at different stages has the potential to allow for timely interventions that may improve a students' learning experience. This paper focuses on two stages: First, **Class Attendance**: Will the newly hired students show up for their first scheduled class? We attempt to predict this at the point of registration. Second, **Training Completion**: Will a student complete all 70 hours of their required training? We attempt to predict this at the point that a student attends his or her first class. As shown in Figure 1, some basic but sometimes incomplete student demographic data are captured at the time a student is assigned to take O&S training. As a student progresses in his or her training journey, we are able to extract more features about learning behavior, such as the amount of time a student needed to complete O&S or the number of days it took a student to register for class. In addition, we leveraged external government census data to augment the existing feature set by adding income and population data of the student's county of residence.

We built four models – Logistic Regression, SVM, Random Forests, and AdaBoost – for the two predicting targets described above. Our final data set contained 5,303 records for predicting first class attendance and 5,182 records for predicting training completion. For both predicting targets, we reserved 2,000 records for testing data set and the remaining were utilized as the training data set. We collected 22 features to predict class completion and used the first 19 features to predict first class attendance(the last three features are not available at our prediction point of registration). Table 1 summarizes the features we used for the model.

3. EXPERIMENT RESULTS

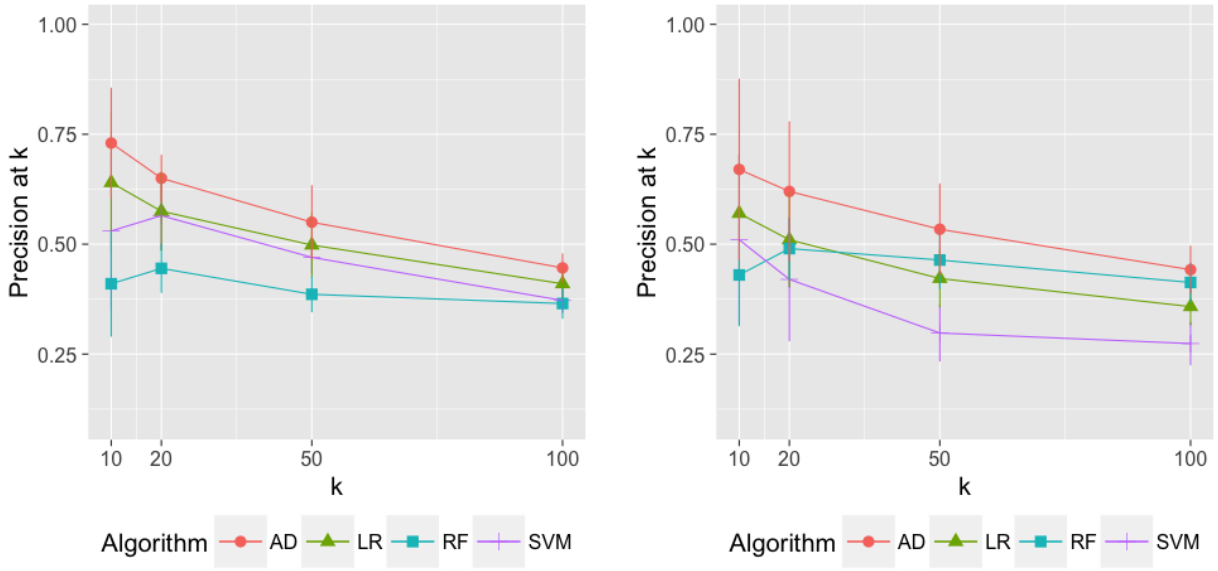
3.1 Prediction Performance: ROC-AUC and Precision at k

We use area under curve of the receiver operating characteristic (ROC_{AUC}) and precision at k (Prec@k) to evaluate prediction quality of each machine learning technique. ROC_{AUC} was used as a standard evaluation metric to measure the quality of overall ranking results. Prec@k was used to determine the quality of predicting the top k outcomes, in our case, the top k students of highest drop out risk at each stage. It is assuming that, with limited resources, front-line staff could only outreach to k number of students per week to provide support and assistance to HCAs struggling to meet their individual learning needs. Therefore, it is essential to accurately predict the first k students exhibiting the highest dropout risk.

Figures 2a and 2b depict the prediction results of our 4 models articulated by precision at k. The AdaBoost model gives the best prediction result for both targets. For predicting first class attendance, AdaBoost with tree number = 2000 has the highest precision at 10 which equals to 0.73 and AdaBoost with tree number = 1000 gives the best precision at 20, 50, 100 which equals to 0.67, 0.56 and 0.46 respectively. For predicting BT completion, AdaBoost with tree number = 100 gives the best precision at 10, 20, 50, 100, which equals to 0.67, 0.62, 0.53, 0.44 respectively. As there are more students who did not attend the first class (385/2000

Table 1: Features used for class attendance and training completion prediction

Feature	Type	Remarks
provider_type	Nominal	Individual provider (paid by the Department of Social and Health Services) or agency provider (paid by private home care agencies). {IP, AP}
student_ethnicity	Nominal	student ethnicity. {Asian Indian, White etc}
student_language	Nominal	student language. {English, Russian, etc}
student_age	Numerical	student age. {Mean = 39, Median = 37}
os_month	Numerical	Month of O&S tracking date. {1,2,$\cdots$,12}
os_day	Numerical	Day of O&S tracking date {1,2,$\cdots$, 31}
class_language_containEnglish	Boolean	Whether the student's profile includes an English language selection. {Yes, No}
class_language_containOther	Boolean	Whether the student's profile includes a language other than English.{Yes, No}
county	Nominal	student's county of residence {King County, Pierce County,etc}
county_income_mean	Numerical	The mean income(in USD) for the county.{mean = 67011, median = 65498}
county_income_median	Numerical	The medium income(in USD) for the county. {mean = 55468, median = 54727}
county_population	Numerical	The population for the county. {mean = 28672, median = 29582}
os_transferredhours	Numerical	Transferred hours for O&S. {mean = 0.9965, median = 0}
duration_to_oscomplete	numerical	Duration(in number of days) beween O&S completion date and O&S tracking date.{mean = 0.842, median = 1.500}
first_module	Nominal	The module of first registered class {Module 1, Module 2,..., Module 20, etc}
duration_to_class	Numerical	Duration(in number of days) between class date and O&S tracking date.{mean=72.05, median = 67.42}
first_class_interpreter	Boolean	Whether the student articulated a need for interpreter services.{Yes,No}
duration_to_class_registration	Numerical	duration(in number of days) between class registration date and O&S tracking date.{mean = 32.647, median = 19.784}
num_terminations	Numerical	Number of terminating employment relationships before attending first class.{0,$\cdots$,7}
student_noshow_count	Numerical	Number of class absences before attending the first class. {0,$\cdots$,58}
student_withdraw_count	Numerical	Number of class withdrawals before attending the first class. {0,$\cdots$,60}
num_class_attendee	Numerical	Number of attendees in the first class. {3,$\cdots$,33}



(a) Precision at k for first class attendance

(b) Precision at k for training completion

Figure 2: Precision at k results

ROC_{AUC}		
Model	1st Class Attendance	Training Completion
SVM(radial)	0.578±0.012	0.600±0.011
LR	0.612±0.020	0.634±0.018
AD(1000)	0.627±0.013	0.673±0.025
AD(2000)	0.626±0.015	0.680±0.024
RF(2000)	0.608±0.012	0.672±0.023

Table 2: ROC_{AUC} results

= 19.25%) than the number of students who did not complete the training (229/2000 = 11.45%), it was slightly easier to predict top k students who were likely to not show up for their first class and explains the higher Prec@k for predicting class attendance.

Table 2 shows ROC_{AUC} results. For predicting first class attendance, AdaBoost with tree number = 1000 gives the best ROC_{AUC} at 0.627. For predicting BT completion, AdaBoost with tree number = 2000 gives the best ROC_{AUC} at 0.68. Low ROC_{AUC} indicates the need for stronger inputs and feature attributes to the models. Although 19 out of 22 attributes were shared in both predicting problems, attributes such as duration to class registration, duration to class and first module were more useful in predicting BT completion than in predicting class attendance. This explains the increased ROC_{AUC} results for BT completion predictions. It provides an opportunity to understand why students choose to not attend their registered training classes and to collect more data at this early stage of the training journey.

3.2 Risk Profile Analysis

In this section, we illustrate how we use insights derived from decision tree modeling to profile students with different dropout rates, providing a tool to isolate target segments of high risk students so the business can take measures that can decrease dropout rate. Decision tree modeling enable us to acquire foundational knowledge necessary to develop educated hypotheses for customized interventions to support students with different risk profiles. Variable importance analysis using Random Forest also enhances our understanding of what factors influence training dropout and assists in our predictions.

At the root note of Figure 3a, the average first class attendance rate is almost 81% among 5,303 students. That is, the overall dropout rate is 19%. For students who didn't enroll in either module 1 or 2 as their first class[1], they demonstrated a significantly higher risk of not attending the training – 54% will not show up for their first registered class. Using the same decision tree, we are also able to infer that both county and age are important factors. For example, students who do not reside in certain counties [2] above and are younger than 49 are less likely to attend the first class compared to those who are older than 49. Younger students, English speaking students and students who take longer to complete O&S exhibit higher risk of not attending their first class. The variable importance from random forest shows that duration to class registration, duration to class are other most important indicators. The larger the time gaps, the higher the dropout rates are.

Figure 3b gives a decision tree for training completion. From the display, we can see if students have two or more class absence records before actually attending the first class, their completion rate decreases to 60%, which is much lower than the average completion rate of 89%. Among these students, if their first class is not Module 1, then the likelihood that the student will complete training drops to 27%. It shows duration to class registration and class location (i.e county) play important role for training completion. Duration to class and student age are also shown as important indicators using random forest variable importance analysis. In addtion, knowing the count of class absence record and first class module gives a much better understanding about the BT completion. Figure 3b shows that even for students who had one or zero class absences. If they register for the class too late (in our case this amounts to more than 52 days after being hired), then the probability of completing the training is even lower.

4. RELATED WORK

Prior studies([3],[7],[8]) have been conducted to explain academic performance and to predict the success or failure across a variety of students in a wide-range of educational settings. These studies focused heavily on the explanatory factors associated with a student's learning behavior and training journey and which of those may cause separation between student types. Machine learning algorithms have been successful in high school and college education settings, most helpful in predicting graduation[4], course participation[5], and other academic outcomes[6]. These algorithms also provide great value to the student success[9].

Lakkaraju et al.[6] used several classification models to identity students at risk of adverse academic outcomes and used precision_at_top_K and recall_at_top_K to predict risk early. The authors compared ROC curves for two cohorts for algorithms Random Forest, AdaBoost, Linear Regression, SVM and Decision Tree. The authors demonstrated that Random Forests outperformed all other methods. Aguiar et al.[10] selected and prioritized students who are at risk of not graduating high school on time by prediction the risk for each grade level and reported precision at top 10%, accuracy, and MAE for ordinal prediction of time to off-track.

[1]Currently, students are allowed to attend classes out of sequence in order to complete their training before the mandatory deadline.

[2]Counties include: Benton, Clark, Cowlitz, Douglas, Grays Hoarbor, Lewis, Mason, Skagit, Stevens, Walla Walla and Whatcom

Johnson et al.[11] used d-year-ahead predictive model to predict on-time graduation for different grade level. Vihavainen et al.[5] found a higher likelihood of failing their mathematics course could be detected in an early stage using Bayesian network. Radcliffe et al.[4] used logit probability model and parametric survival models to found that demographic info, academic preparation and first-term academic performance have a strong impact to graduation. Dekker et al.[12] gave experimental results which showed decision trees gave a high accuracy for predicting student success and improved prediction accuracy using cost-sensitive learning.

Other prior studies have highlighted some important indicators that influence students' performance like a student's age and absence rates[6]. Based on these features, Early Warning Indicator (EWI) systems are rapidly being built and deployed using machine learning algorithms[6]. Similar to other research in Educational Data Mining (EDM), we use precision at k to measure the prediction result([6], [10], [13]) and, like in traditional education systems, our motive is to most effectively and efficiently target our limited resources to assist and suppor students. Typically, ensemble models outperformed individual models[7] and this held true in our case as well. While random forest has proven to be an extremely useful and powerful machine learning technique in educational research[11], our results indicated that AdaBoost outperformed random forest.

5. CONCLUSION AND FUTURE WORK

In this study, we demonstrated preliminary results for predicting home care student training dropout from a large, heterogeneous dataset containing student demographics and engineered features extracted from training patterns. Predicting dropout at varying stages of an adult learner's training journey yielded promising results from a skewed dataset of over 5,303 students with AdaBoost (2,000 trees) providing the strongest predictions (prec@10 = 0.73 and ROC_{AUC} = 0.625. Prior history of class absence and time effects (duration to registration, duration to first class) were among the strongest individual predictors of dropout, as were class module sequence, county, and student age. The results demonstrate that applying machine learning techniques to demographic data and learning behavior data (e.g. duration to registration, duration to first class) can achieve adequate prediction quality in predicting the top k highest risk students out of a pool of newly hired HCAs. This enables efficient use of limited capacity and resources to support students of greatest need. Insights revealed in this study inspired training operation staff to explore alternatives, including encouraging newly hired HCAs to register for training early and strongly recommend proper class sequence to support students success in their training.

Future work will investigate collecting more information about students, such as their motivations, propensity for self-efficacy, and life circumstances to determine if there are other factors at play on a personal level that my uncover additional features that can contribute to our target predictions around training dropout.

6. REFERENCES

[1] Charissa Raynor. Innovations in training and promoting the direct care workforce. *Public Policy & Aging Report*, 24(2):70–72, 2014.

[2] Colombo Francesca, Llena-Nozal Ana, Mercier Jérôme, and Tjadens Frits. *OECD Health Policy Studies Help Wanted? Providing and Paying for Long-Term Care: Providing and Paying for Long-Term Care*, volume 2011. OECD Publishing, 2011.

[3] S Kotsiantis, Christos Pierrakeas, and P Pintelas. Predicting students'performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5):411–426, 2004.

[4] P Radcliffe, R Huesman, and John Kellogg. Modeling the incidence and timing of student attrition: A survival analysis approach to retention analysis. In *annual meeting of the Association for Institutional Research in the Upper Midwest (AIRUM)*, 2006.

[5] Arto Vihavainen, Matti Luukkainen, and Jaakko Kurhila. Using students' programming behavior to predict success in an introductory mathematics course. In *Educational Data Mining 2013*, 2013.

[6] 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.

[7] Dursun Delen. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506, 2010.

[8] S Kotsiantis, Kiriakos Patriarcheas, and M Xenos. A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, 23(6):529–535, 2010.

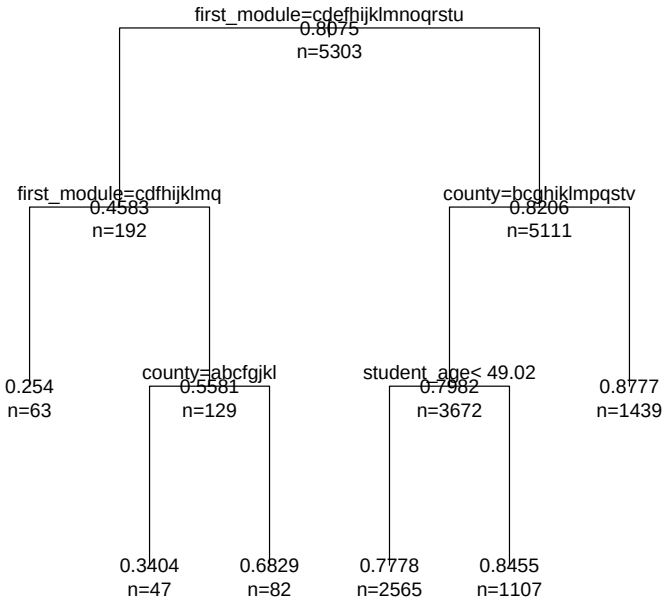
[9] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[10] Everaldo Aguiar, Himabindu Lakkaraju, Nasir Bhanpuri, David Miller, Ben Yuhas, and Kecia L Addison. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 93–102. ACM, 2015.

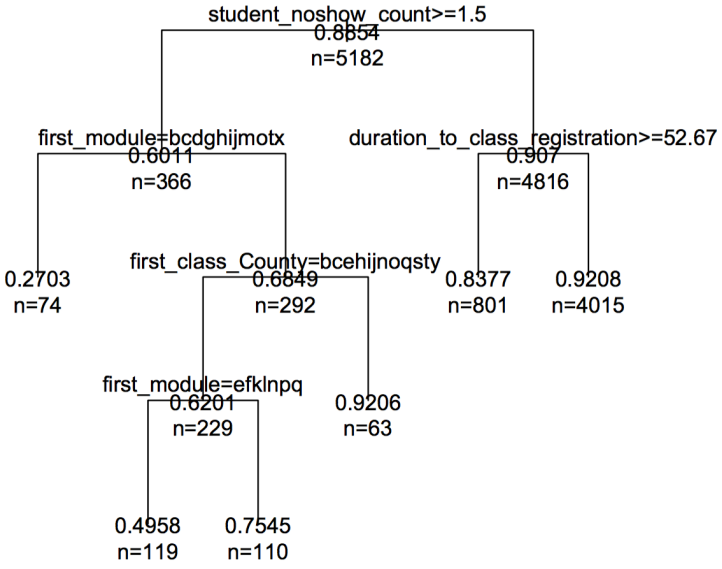
[11] Reid A Johnson, Ruobin Gong, Siobhan Greatorex-Voith, Anushka Anand, and Alan Fritzler. A data-driven framework for identifying high school students at risk of not graduating on time.

[12] Gerben Dekker, Mykola Pechenizkiy, and Jan Vleeshouwers. Predicting students drop out: A case study. In *Educational Data Mining 2009*, 2009.

[13] Everaldo Aguiar, G Alex Ambrose, Nitesh V Chawla, Victoria Goodrich, and Jay Brockman. Engagement vs performance: Using electronic portfolios to predict first semester engineering student persistence. *Journal of Learning Analytics*, 1(3):7–33, 2014.



(a) First Attend



(b) Training Completion

Figure 3: Decision Trees