Using Modern Statistical Methods to Analyze Demographics of Kansas ABE/GED®
Students who Transition to Community or Technical College Programs

Abstract

This research analyzed linked high-quality state data from K-12, adult education, and postsecondary state datasets in order to better understand the association between student demographics and successful completion of a postsecondary program. Due to the relatively small sample size compared to the large number of features, we analyzed the data with Nearest Shrunken Centroid (PAM), a statistical method developed for cancer and genomic research. Our findings conclude that there are five features that are predictive of an adult education student's successful completion of a postsecondary program.

Keywords: ABE/GED, Student Pathways, Nearest Shrunken Centroid, Linked Datasets, Postsecondary Transition

Introduction

The goal of this research was to analyze the high-quality state data that linked the Kansas K-12, adult education, and postsecondary datasets in order to better understand the association between adult education student demographics and successful completion of postsecondary programs. We have two well-defined classes (completers vs. non-completers). The purpose is to identify covariates that are predictive of the given class. Such data are typically analyzed with a classification method, in which there are explanatory variables and response variable, and the purpose is to identify contributions of different explanatory variables and predict the response variable. Commonly used classical methods such as logistic regression, CART, linear discriminant analysis, and quadratic discriminant analysis have limitations due to curse of dimensionality, i.e., a relatively small sample size with an extremely large number of features (potential predictors). In this research a statistical method, Nearest Shrunken Centroid (PAM), was used to identify student demographic and coursework information that were associated with student success. PAM (also referred to Prediction Analysis of Microarrays) was developed for genomic and cancer research to analyze small samples with a large number of features. PAM can handle multi-class as well as two-class cases. In Tan, Naiman, Xu, Winslo, and Geman (2005), PAM was applied to 9 two-class and 10 multi-class cancer genomics datasets. The advantages of using PAM are: (1) It is particularly developed for feature selection and classification in datasets with high dimensional explanatory variables but limited sample size, in which case frequently used classical methods such as logistic regression, CART, linear discriminant analysis, and quadratic discriminant analysis face the limitation due to curse of dimensionality; (2) PAM is extremely computationally efficient because it does variable selection through thresholding that gets rid of variables with small effects easily in a single step for each thresholding parameter value; and (3) PAM has high classification accuracy because it aims at maximizing the posterior probability of belonging to a class given the observed data. It is particularly powerful when the explanatory variables are continuous. Even with binary explanatory variables, the performance of PAM is surprisingly good and robust compared to other methods tailored for high dimensional classification problems.

The significance of this study is that specific sets of features that associate with completion were identified for students who were enrolled in a public high school, transitioned to an adult learning center, earned their GED® credential, and transitioned to a postsecondary institution. It is also significant because this research used a statistical method not normally used in education research, and has potential to be used with other adult student datasets that have small sample size and large numbers of features.

Research Problem

This research used Kansas' linked student data that was collected by secondary, adult education, and postsecondary schools between 2007 and 2012. Each case in the dataset represents a student who attended both a Kansas public secondary school and an adult education center, and then transitioned into a Kansas postsecondary program. In the field of adult education, little quantitative research exists on demographic-specific patterns of adult education students who successfully transition to postsecondary education after earning a GED® credential and factors.

This research systematically explores this linked dataset to determine if demographic factors and coursework patterns are associated with student pathways from secondary education through adult education to postsecondary education. This type of research has the potential to identify adult basic education (ABE), adult secondary education (ASE), and English as a second language (ESL) student factors that increase the odds of successfully completing a postsecondary program. Findings can be used to develop policies and design interventions that increase student achievement and help state adult education programs achieve their goals by exploring sensitivities to factors. Our research contributes to the national literature on adult education students' pathways into postsecondary education linked to student academic achievement. What we have learned suggests that the pathways of these adult education students are, in fact, quite complex.

Background

Nationally, there is great interest in identifying the education pathways that adults follow when transitioning from secondary to postsecondary education, and both public and private-funded research are currently examining this issue. One such investment is a two-year project conducted by the National Center for Higher Education Management Systems (NCHEMS) with funding from the Bill and Melinda Gates Foundation, which examined state policies that foster student progression and success in the 'adult re-entry pipeline' (Boeke & Zis 2011; Ewell, Kelly, & Klein-Collins, 2008). In 2007, the U.S. Department of Education, Office of Vocational and Adult Education (OVAE is now the Office of Career, Technical and Adult Education or OCTAE), also signaled its continued commitment to address this issue by awarding four grants through the *Ready for College: Adult Transitions Programs* to implement projects focused on improving the quality of ASE, so that out-of-school youth can successfully transition to postsecondary education.

OCTAE is especially interested in what can be learned from state adult education databases, and requires that each state keep data on those students who successfully transition to postsecondary programs. In the Adult Education Annual Report to Congress (U.S. Department of Education, Office of Vocational and Adult Education, 2007), data from OVAE show that only one-third of adult education students with a documented goal to enroll in postsecondary education or training actually transition within an academic year period. Bragg, Kim, and Barnett (2006) note that over 90 percent of high school sophomores self-report that they want to go to college and 70 percent want to complete a four-year degree (p. 5). For many, these dreams are never realized. According to the U.S. Department of Education, the *status dropout* rate in 2009 for 16-24 year olds who have not graduated or earned a GED® credential was overall about 8 percent, with 5 percent for Whites, 9 percent for Blacks, and 18 percent for Hispanics

(IES-NCES, 2012, p. 240). In 2009, about 75 percent of the high school students graduated on time, and in Kansas the rate was 80 percent (IES-NCES, 2012, p. 268). The end result is that approximately 6,000 Kansas high school students each year do not graduate on time, and many enter one of Kansas' adult education centers to pursue a GED® credential.

Of those who take the GED® test nationally, an estimated 60 percent indicate that they do so in order to begin postsecondary education. However, not all GED® passers follow up on their intentions. Data matched from two cohorts (consisting of a half million students each from 2003 and 2004) of the GED® test-taking population with National Student Clearinghouse postsecondary data indicate that nearly 43 percent of GED® passers enroll in postsecondary education, yet only 12 percent of enrollees graduate within six years (Patterson, Zhang, Song, & Guison-Dowdy, 2010; Zhang, Guison-Dowdy, Patterson, & Song, 2011). Typically adults with GED® credentials enroll in a community college, at least initially, to continue their education (Reder, 2007).

The problem of persistence and retention as factors of student success in higher education has been studied for decades (Acee, Cho, Kim, & Weinstein, 2012; Bean & Metzner, 1985; Pascarella, 1985). These studies tend to focus on student integration, social support systems, personal commitment, high school achievement, social economic conditions, and institutional practices. Horn (2014) delineates five broad interrelated factors that determine student retention and completion in postsecondary education: institutional practices, social identification, goal commitment, academic engagement, and student success. Institutional practices include academic advising, faculty engagement, financial aid, remedial education, and student assessment. Social identification includes internalization of student norms, sense of belonging, and quality of relationships with teachers and peers. Goal commitment includes postsecondary credential, time to completion and level of vocational and civic engagement. Academic engagement includes time on task, deep conceptual learning, and interest and enjoyment. And student success includes academic achievement, and persistence and completion (p. 3). Miller's (2014) synthesis of graduation rates for nontraditional students posits that there are multiple factors affecting adult student postsecondary completion. His study indicates that student completion rates are affected by degree type and enrollment intensity where certificates have higher completion rates than associate or bachelor degrees, and fulltime students graduate at higher rates than part-time students.

Prior research on adult student pathways, for the most part, has been descriptive, often simplified, and does not associated demographic and other factors with various pathways through adult education programs to postsecondary education. Nor has this prior research been used to investigate implications for adult education interventions. For example, our dataset connects student information with the adult education center they attended. If location is predictive of success, we can investigate practices, curricula, and other factors around program effectiveness in high and low performing adult education centers, and then replicate what works in high performing education centers in low performing centers.

Kansas Background

In Kansas, members of the 2003 cohort enrolled in community colleges at a rate of nearly 30 percent, with 11 percent graduating from their postsecondary program in six years (Patterson, 2010). The 2004 Kansas cohort members enrolled in postsecondary programs (at all levels) at a rate of 41.8 percent, and 10 percent graduated within six years (Research Allies for Lifelong Learning, 2013). While assisting adult education students with their postsecondary education transition goals has been a component of Kansas adult education programs since the 1998

Workforce Investment Act legislation, concentrated transition efforts and performance-based funding for Kansas adult education programs did not begin in earnest until 2005, which suggests that these percentages likely underrepresent the actual percentages for postsecondary transition outcomes of Kansas adult education students.

In 2010, the Kansas Board of Regents (KBOR) adopted a 10-year strategic plan, *Foresight 2020,* which set long-range achievement goals for all of the entities "to ensure the state's higher education system meets Kansans' expectations", including all of state's adult education centers (p. 1). Presently, Kansas' adult education programs serve 8,000 to 9,000 students each year, out of 230,000 Kansans who are eligible adults (Developmental Education Working Group, 2014, p. 25). In order to meet *Foresight 2020* goals with limited state and federal funds, there is a need to better understand which type of student, based on individual and institutional demographics, is more likely to succeed in an adult education program and transition into a postsecondary education program.

We also know that it is possible for adult education students to enter a postsecondary program without having earned their GED® credential. Many of these students never enroll in an adult education center. All Kansas community and technical colleges have open enrollment so adult education students can enroll without having a high school diploma or GED® credential. Yet at the same time we know that GED® students do not do as well in terms of persistence and graduation as high school graduates in postsecondary programs (Guison-Dowdy & Patterson, 2011; 2011a). Many high school dropouts who enroll in a community or technical college are placed in developmental education courses, which still require tuition and extend the time to completion. Moreover students who enter developmental education are less likely to complete a postsecondary program (Bailey, 2009; 2009a). There are multiple pathways for students entering postsecondary, many of which bypass a high school credential. The percentage of students who successfully enroll in and complete a postsecondary program is significantly lower for those who follow such nontraditional pathways.

Method

The goal of this research was to identify predictive factors of an adult education student who earn a GED® credential and successfully complete a postsecondary program. Some factors may be malleable where an intervention can be designed to strengthen student achievement. For example, if years completed in high school is predictive of future success in adult education and postsecondary education, then an intervention can be developed to keep at-risk students in high school as long as possible.

Successful completion of a postsecondary program is only counted in our study if it occurs between 2007 and 2012. Hence if a student is still making progress in 2012 but has not completed a postsecondary program they are counted as a non-completer. Postsecondary education in this research refers to any education program at a technical college, community college, college or university, but does not include adult education programs hosted by technical and community college.

Dataset

All students in our dataset met these requirements: between 2007 and 2012 they attended a Kansas public high school, enrolled in a Kansas public adult learning center, and transitioned into Kansas public postsecondary school. The total number of student records that met these criteria was 2,258. This total included duplicate cases where the same student had multiple records across different years, in multiple adult education centers, or multiple postsecondary institutions in multiple years. We removed all duplicate cases, as we were only interested in the final year they were enrolled in an adult education center, the first year in a postsecondary program, and whether or not they completed a postsecondary program, which reduced our dataset to 532 unique cases (students).

KBOR has oversight of all public adult education centers, technical and community colleges, four-year colleges, and universities. KBOR also has a data-sharing agreement with the Kansas State Department. This administrative structure and interagency agreement allowed us to use linked student data from high school, adult education, and postsecondary education. In addition, all Kansas public adult education centers use the same data collection and testing protocol, including several audits and quality assurance steps. As a result the Kansas' adult education dataset is of very high quality.

The demographic composition of 532 cases of our sample is mostly young adults, predominantly white, with slightly more females than males.

Table 1: Student Demographics

Gender	%
Male	47
Female	53
Age	%
16-19	59
20-24	38
25+	3
Ethnicity	%
African American	8
Hispanic	10
White	70
All Other	12

For each student our dataset contains these variables:

- Whether the student enrolled in developmental education course at a community or technical college, including development English, reading, math and any developmental course;

- Whether the student declared a major in a postsecondary program. This variable has five categories: declared major in associate degree program, declared major in stand-alone program (a certificate requiring less than 16 credit hours), declared major in a longer certificate program, declared major in a STEM program, and no declared major;
- If the student earned a GED® credential;
- If the student completed high school diploma and also enrolled in an adult education program;
- If the student received public assistance—self reported;
- If the student received a Pell Grant;
- The student's gender;
- The student's race—self reported;
- The student's age;
- The last fiscal year the student was in an adult education program;
- Total hours the student was in an adult education program;
- Each adult education center's institutional ID;
- Each technical and community college institutional ID;
- The student's first postsecondary year;
- The student's time in post-secondary program; and
- The student's pass rate (there are three levels of pass rate: enrolled but did not pass any courses in a postsecondary program, passed at least one course but less than 50 percent of their courses, and passed 50 percent or more of their courses).

We are particularly interested in the main effect and two and three-way interactions of these variables, and how the interactions are associated with completing a postsecondary program.

Statistical Method

The main effects (single factor), two-way, and three-way interactions of all main-effect variables were included in our analysis. This leads to a model with 26,534 potential predictors, which we refer to as features in this paper. To estimate the effects of such a large number of predictors in a classic regression model, the sample size needs to be much greater than the number of features, generally a cases-to-predictors (n to p) ratio of 10:1 or more is required Since the sample size is 532, which is much smaller than the total degrees of freedom of all the variables and their interactions, classical regression models fail due to insufficient degrees of freedom in calculating the sum of squares of error. In this analysis, we employed a modern method that had been developed for high-dimensional data, in which the number of predictors can be much higher than the sample size.

The Nearest Shrunken Centroids method was used to analyze the data. It is one of the most popular classifiers employed in cancer classification problems using microarray data from genomics. It is more commonly known as the Prediction Analysis of Microarrays (PAM). It can work with multi-class and two-class systems where there can be multiple or two outcomes and each outcome comprises a class. In our study there were two classes, one for failing to complete any program (class 1), and the other one for successfully completing one program (class 2). A standard *centroid* for each class is computed for each feature, which is the average of the feature values for students in the same class, divided by the within-class standard deviation for that feature. The feature profile of each student is compared to the standard centroid to classify cases. The PAM identifies important features by shrinking each centroid toward the overall centroid and removing features below the threshold. Cross-validations were used to choose the threshold parameter of the amount of shrinkage that minimizes the classification error. PAM is powerful for two reasons: it uses thresholding to reduce noise, and the original classifier before achieving the shrunken centroids is the Naive Bayes classifier. This original classifier assumes that the conditional distribution of feature variable given in a class is normally distributed with feature specific variance. If the conditional distribution of feature values given students in each class is correctly specified, the Naive classifier gives the optimal classification according to Bayes Theorem. PAM can perform predictor selection and feature selection in its modeling process. Feature selection and class modeling were conducted iteratively.

One potential limitation of the PAM process is the assumption that the predictors follow a normal distribution within the classes. However, violation of this assumption in the sparse direction (for example, when there are a lot of zero values in the dataset) does not pose a threat to its classification accuracy. An example of such robustness to violation of the normality assumption in the sparse direction is the application of PAM on the Leukemia3 dataset in Tan, et al. (2005), which contained 7 cancer classes and 12,558 feature variables. More than 65 percent of the feature values were zero. Even in such an extreme situation, PAM gave the highest classification accuracy of 93.75 percent among all tested methods (Tan et al., 2005; Wang, Dai, Chen &Yuan, 2013). In our dataset many of the feature values are also zero. All the analyses in this study used R version 3.0.1.

Data Analysis

We conducted two types of analyses. The first type was an external 10-fold cross-validation. The subjects in each class of the entire dataset were randomly partitioned into 10 subsets. Then the subjects in class 1 and those in class 2 from each subset were combined to yield a single subset that includes both students from class 1 and students from class 2 with the same ratio as the original dataset. In the 10-fold cross-validation, there were 10 model fittings and 10 predictions. This validation was an iterative process in which each subset is left aside as a test set and remaining subsets were used as the training set to build the model. The obtained model was then validated on the test set to assess the prediction performance of the model. The process continued until every subset had been used as the test data. In the end, the predicted classifications from all subsets were compared to the observed classifications to assess the proportion of correctly classified students by their program completion status. In this external cross-validation, the feature selection and model fitting were performed 10 times so that there would be 10 models and different sets of selected features for different training sets. Since the model fitting did not use any data from the test set, the resulting reported accuracy is representative of the performance for other independently observed data. This cross-validation procedure establishes the validity of the model.

The second type of analysis was to conduct model fitting (and feature selection) with the entire dataset. Afterward, the selected features and model were used in 10-fold cross-validation. Even though there was classification accuracy reported from this analysis, the accuracy is not generalizable to future datasets since all subjects have been used during the feature selection and model fitting process. The PAM method selects features but does not provide *p*-values as in classical regression analysis. After the features were selected, we ran a logistic regression model. If we use the default cutoff of 0.5 on the posterior probability to determine the predicted class, the overall accuracy of the logistic regression model was high. Unfortunately, it did not correctly classify any of the program completers. When different cutoff values were used on the posterior probability as shown in a ROC curve, the logistic regression and the Firth model using the features selected by PAM produce similar area under the ROC curveas the PAM. Without the feature selection step from PAM, neither the logistic regression nor the Firth logistic model could provide best fit to our data that have relatively small sample size and a dichotomous outcome variable with a highly disproportionate distribution in additional to high dimensional explanatory variables. Therefore, PAM can be used as a feature selection tool prior to an application of logistic or Firth logistic regression model.

Findings

We first consider the results obtained by setting the objective function to be the overall classification accuracy in order to find the optimal threshold parameter(s) and select predictors using the training data. In this case the overall percent of correctly classified samples is maximized over all possible feature subsets achieved by using different values of the threshold parameter. Table 2 reports the number of samples classified into each class as well as classification accuracy as a percentage within each class.

Table 2: PAM Prediction of Classes

	PAM Prediction		
	Class 1-predicted	Class 2-predicted	Percent Accuracy
Class 1-actual	418	44	90.48
Class 2-actual	27	43	61.43
Overall Percent			86.65
MCC			0.474

Cross-tabulated counts of the observed sample class with the predicted class from 10-fold external cross-validation using classification accuracy as the objective function during the model training. Observed: 462 in class 1 (fail to complete any program) and 70 in class 2 (successfully complete a program). The percent refers to the percent of correctly classified samples in a row. MCC is the Matthews Correlation Coefficient.

Table 2 shows that the PAM model better predicts who will not complete any program than who will successfully complete a program. Non-completion is predicted correctly in 418 of 462 cases (90.48 percent accuracy), but completion is only predicted correctly in 43 of 70 cases (61.43 percent accuracy).

The distribution of the outcome variable is highly unbalanced. In such case, maximizing the overall classification accuracy may not accurately predict who completes a program. A better performance measure on unbalanced datasets is the Matthews Correlation Coefficient (MCC) defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}$$

where TP is the number of true positives (i.e. correctly classified students who completed a program), TN is the number of true negatives (i.e. correctly classified students who did not complete a program), FP is the number of false positives (i.e. students who did not complete a program but were misclassified) and FN is the number of false negatives (i.e. students who completed a program but were misclassified). The values of MCC are in the range of [-1, 1] with value 1 indicating a perfect prediction and -1 an extremely opposite prediction. It should be noted that the thresholding parameter in PAM was estimated by minimizing the 5-fold cross-validation accuracy within the training data. It is possible to change the objective function to replace the 5-fold cross-validation accuracy by cross-validation MCC.

Table 3 gives the result of PAM using MCC as the objective function during the model training and feature selection process.

Table 3: PAM Prediction using MCC

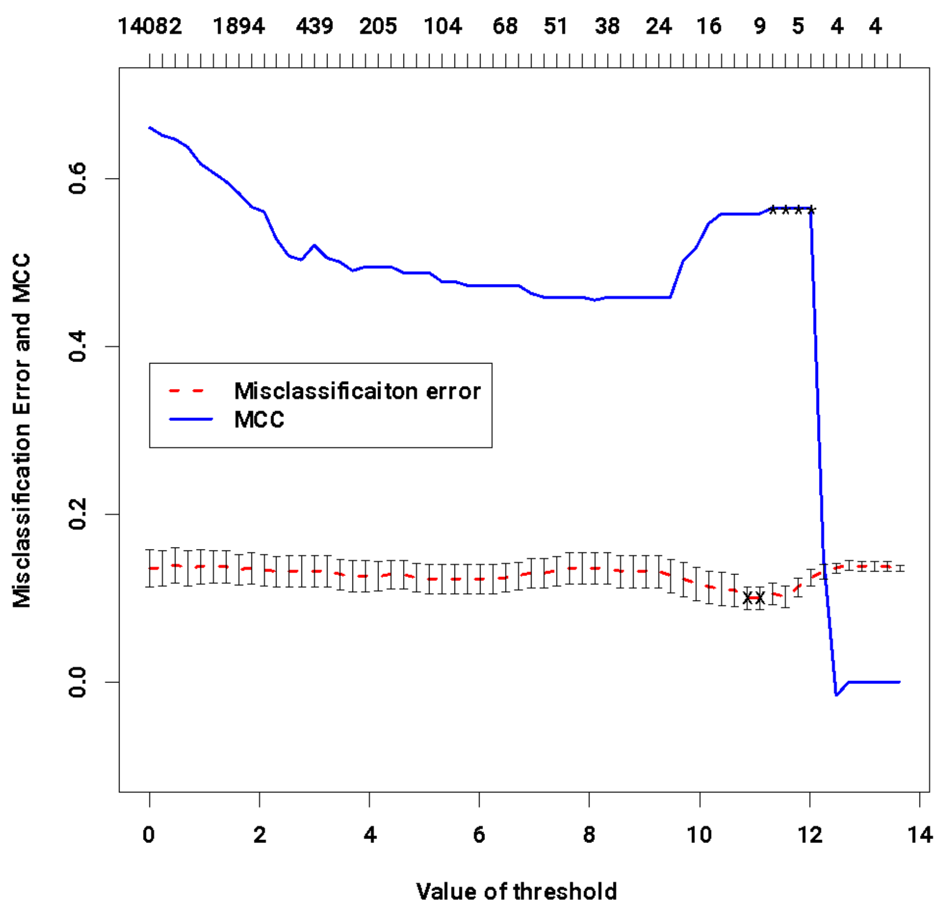
	PAM Prediction		
	Class 1-predicted	Class 2-predicted	Percent Accuracy
Class 1-actual	432	30	93.51
Class 2-actual	30	40	57.1
Overall Percent			88.72
MCC			0.506

Cross tabulated counts of the observed sample class with the predicted class from 10-fold external cross-validation using MCC as the objective function during training. Table legend is same as in Table 2.

It can be seen in Table 3 that 432 of 462 non-completers (93.51%) are correctly classified and 40 of 70 completers (57.1%) are correctly classified. Compared to the results in Table 2, a summary from the external cross-validation is that using MCC as the objective function yield higher overall accuracy and MCC values.

Next we report the PAM results from fitting models with parameters and number of predictors selected using the entire dataset. Figure 1 shows the variation in the MCC as the threshold value is increased.

Figure 1. PAM Misclassification error and MCC vs. threshold value



The horizontal axis on top labels the number of features selected as the threshold parameter was given in the horizontal axis at the bottom. The '*' labels the MCC of models that has the highest MCC values among models using less than 1000 predictors.

The highest MCC was achieved initially with 14,082 features and then the MCC was reduced as the value of the thresholding parameter increased, and then MCC increased again for thresholding parameter values between 10 and 12. Increasing the thresholding parameter corresponds to reducing the number of selected features. The objective in creating this model is to maximize the MCC and obtain a theoretically interpretable model. For interpretability reasons, the maximum number of features was limited to 1000. Figure 1 shows that the maximum value of the MCC and an acceptable number of features (< 1000) occur at a threshold of 11.33; the same value of MCC occurred for multiple values of the threshold.

The models of the predictors of program completion created using PAM, with the best MCC, show that students with interactions of the following variables were more likely to succeed: declared major in a stand-alone program, improved their pass rate from one level to the next, did not enroll in developmental education, the first postsecondary year enrolled is 2011, and were older. Hence, interactions of these variables result in the highest likelihood of completing a program in order of importance as follows (: indicates interaction of the two variables):

- Declared major in a stand-alone program : pass rate : age
- Declared major in a stand-alone program : pass rate
- No developmental education : declared major in a stand-alone program : pass rate
- Declared major in a stand-alone program : pass rate : no developmental math
- Declared major in a stand-alone program : postsecondary year 2011: pass rate

To check for potential bias in the results from our PAM model, we used Firth's logistic regression model, which is a penalized likelihood approach to reduce small-sample bias in maximum likelihood estimation of logistic model with rare events. In our case, we have 70 completers and 462 non-completers. The 'completers' are therefore rare events. Even though the Firth's model is meant to reduce bias for small sample size, it was developed for traditional settings where the number of features is not overwhelmingly high. We have 26,534 features with the sample size of 532. It is not possible to complete model fitting or variable selection using Firth's model. An application of Firth's model to a reduced dataset that contains only some selected features would be helpful to increase the prediction accuracy of the completers. PAM does a good job of feature selection. Surprisingly, the Firth model with the features selected by PAM gives even better prediction results than PAM itself (in terms of prediction for the completers). Therefore, a combined application of PAM doing feature selection and Firth model doing prediction would be more beneficial than applying either method alone.

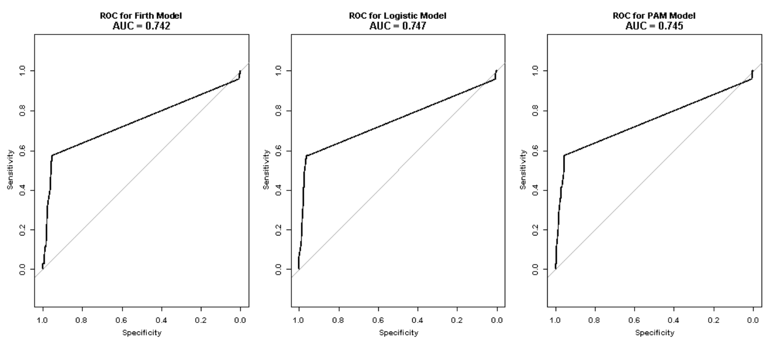
Table 4: Firth Prediction with Features Selected by PAM

	Firth Prediction **with probability cut-off of 0.5**		
	Class 1-predicted	Class 2-predicted	Percent Accuracy
Class 1-actual	448	14	96.97
Class 2-actual	46	24	65.71
Overall Percent			88.45

Cross tabulated counts using the Firth logistic regression model with five features selected by PAM. Table legend is same as in Table 2.

We then plotted the ROC curves using the fitted probability of completers with the models using the selected features. The ROC curves report the true positive rate (vertical axis) and false positive rate (1- horizontal axis) at varying levels of cut-off threshold for the posterior probability. The plots are shown in Figure 2. The left-most panel is from Firth logistic regression model, the middle panel is from simple logistic model, and the right-most panel is from the PAM model. All three models produced nearly identical area under the curve (AUC). This indicates that the features selected by PAM are not limited to have predictive power with PAM but also have good predictive power for logistic regression or Firth logistic regression model.

Figure 2: ROC curves of the fitted Firth logistic regression, simple logistic regression, and PAM models using the selected features from PAM.



Discussion

Though it is tempting to attribute success to one variable or a main effect, the features that are most predictive are two- and three-way interactions. Analyzing each individual variable or main effect feature may provide a glimpse as to why these interactions are probable predictors but in this analysis no individual feature was predictive of an adult education student successfully completing a postsecondary program. “Declaring a major in a stand-alone program” and “increasing pass rate” is in each of the five features that were most predictive of a student successfully completing a postsecondary program. Not taking developmental education was in two of these features, and age and entering postsecondary in 2011 were in one feature.

Declaring a major in a stand-alone program (a certificate requiring less than 16 credit hours) when interacting with other features appears to enhance a student's odds of completing a postsecondary program. This result might be explained by the shorter required time commitment and a higher perception of goal attainability. If declaring a major is an indicator of clear sense of goals, this may be suggestive of other positive factors such stronger motivation, greater clarity of purpose, more maturity, more confidence, and higher level of comfort in formal schooling. It could also be due to the fact that most of these short-duration programs are more technical than academic, therefore not requiring higher levels of reading, writing and mathematics, and that most of these programs are designed as career programs where the curriculum is tied to job skills. More research is needed to better understand why a declared major in a longer program, such as certificate or associates degree program when compared to a stand-alone program, was not identified in our research as a factor to increase the likelihood to complete a postsecondary program.

Increasing student pass rate in a postsecondary program was also in all the interactions that increased a student's odds of successfully completing a postsecondary program. It can be interpreted as a sign of persistence and retention, as well as higher academic ability. The policy implication is that keeping a student in a postsecondary program where they continually make progress is more likely to associate with successful completion than dropping all courses in any given semester.

Not enrolling in developmental education was in two of the interactions, which concurs with Bailey's (2009) findings that only 44 percent of students who enrolled in developmental education courses completed the recommended remedial courses within three years (p. 14). The problems with developmental courses are that they do not lead to a degree, they burden students with additional tuition costs, and they further delay a student's career and education goals. Not enrolling in developmental education courses might also be associated with having higher ability and being more prepared for a postsecondary program.

First year of postsecondary year in 2011 and increase in age are only in one predicted feature. When comparing the year of entry in a postsecondary program, we discovered that in 2011 cohort had a greater likelihood to succeed when compared to other years when those individuals declared a major in a stand-alone program and improved their pass rate. This may represent a research limitation as our study classified students as non-completers even though they were still enrolled in 2012. Also, since 2011 cohort occurs near the end of our study period (2007-2012) its significance may be accounted for via improved advising, counseling or other student support structures given the state's efforts to improve retention in Kansas technical and community colleges. Age when aligned with declaring a major and pass rate may logically indicate that older students have clearer goals and stronger levels of persistence and motivation even though 97 percent of the students in our dataset are younger than 25 years. Both of these main effects are only significant when associated with other factors, hence to understand their level of significance requires more research.

While it may seem counterintuitive that single factors (main effects) are not predictive but their interactions are, these interactions do indicate the complexity of why students go down one pathway instead of another. If a single factor were a significant predictor of completing a postsecondary program, PAM would have identified it. Our research therefore raises questions regarding the value of using single-factor analyses that do not consider interactions to understand events. Similar to understanding factors that determine what causes cancer, the reasons why someone succeeds or fails in education is equally complex and with the advancement of new statistical methods we now have the tools to better understand the interactions of main effect factors in postsecondary completion.

Limitations

There are several limitations to this research. Though our dataset is of high quality, our findings are still dependent upon the available variables of this dataset that links secondary, adult education, and postsecondary data. We also were limited to employing the existing demographic features in this database some of which is self-reported. Hence we cannot address the influence of advising, a sense of wellbeing, inaccurate self-reports, maturity, and other factors that may predict student success in postsecondary programs. Another limitation is that to our knowledge this is the first study to analyze student demographics and coursework using these methods, hence we have no similar studies with which to compare our study. And this research is limited to only those students who met the criteria of being in a Kansas secondary school, adult education program, and postsecondary program between 2007 and 2012. This time frame limits the age range to younger adult education students, counts a student as a non-completer even though they are still enrolled and making progress after 2012, and does not account for those students who bypass an adult education program or gain a high school credential by passing the GED® test without instruction. As we were interested only in completers and non-completers, further research using PAM is needed on non-completers were still enrolled after 2012.

This study did not include subjects that had multiple records covering more than one year. If these subjects were included, correlation among records from the same subject needs to be considered during the modeling. This would incur problems in two regards. (1) Majority of subjects only have records for one year and only a small portion of subjects have records for more than a year. Including both into the same model makes it difficult to model the correlation appropriately since subjects with only a one-year record does not contribute to the estimation of correlation structure. (2) So far the study of methods with correlated data and high dimensional explanatory variables is still an active research topic. There are some theoretical results exist but we have not seen any methods that have general applications.

Conclusion

Since 2007 in Kansas, between 1,700 and 2,200 students annually pass the GED® test and earn their Kansas State High School Diploma. It is important to note that not all of these students pass through an adult education program. Also approximately 1,000 adult education students enroll annually in postsecondary education or training programs (OVAE-National Reporting System, Table 5), though not all earn a GED® credential. Based on what we know, the majority of students who earn a GED® credential enroll in Kansas' community and technical colleges. And, increasing the numbers of these students who successfully transition to and complete a postsecondary program is closely aligned with KBOR's goal to double the number of adult education students entering postsecondary programs after attending adult education. This goal reinforces why analyzing state data is critical to better understand demographic-specific patterns and factors of adult education students who successfully transition to postsecondary programs.

Our findings do not show that individual factors were predictive in the five predictive interactions; rather, it is the five two- and three-way interactions that were predictive. These interactions illustrate the complexity of understanding student pathways between adult education and postsecondary programs. In order to capture this complexity we had to use a statistical method (PAM) developed for cancer research that is designed to analyze relatively small samples with large numbers of features. Higher quality national and state student databases present an important research opportunity to facilitate better understanding of the predictive factors in student success. Often these datasets contain many demographic features relative to the size of the sample. For example the USA-PIAAC (Program for International Assessment of Adult Competencies) dataset contains many features for a relatively small data sample, 5,000 adults (IES-NCES, 2012a). To isolate one or two features to explain or predict an effect using a classical model does not take into consideration the interaction with other features. The methods we used identified complex interactions associated with adult pathways and success.

Our research findings also provide insight into how linked SLDS datasets can be used to study student achievement, as well as teacher and program effectiveness. Nationally there is a movement to increase student transitions to postsecondary education and identify the best models for doing so, with a focus on certain subpopulations, namely adult education students. The type of analyses used in this study has the potential to be applied to other state adult student datasets, and further our understanding of why some students passing through adult education programs succeed or fail in postsecondary programs.

References

Acee, T. W., Cho, Y., Kim, J. I., & Weinstein, C. E. (2012). Relationships among properties of college students' self-set academic goals and academic achievement. *Educational Psychology, 32*(6), 681-698.

Bailey, T. (2009). Challenge and opportunity: Rethinking the role and function of developmental education in community colleges. *New Direction for Community Colleges, 2009*(145), 11-30.

Bailey, T. (2009a). *Rethinking developmental education in community college*. New York: Community College Research Center.

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55*(4), 485-540.

Boeke, M, & Zis, S. (2012). *State policies on the adult re-entry pipeline state case studies.* Boulder, CO: National Center for Higher Education Management Systems. Retrieved from http://www.nchems.org/c2sp/documents/GatesAdultCaseStudies_Final.pdf.

Bragg, D. D., Kim, E. & Barnett, E. A. (2006). Creating access and success: Academic pathways reaching underserved students. *New Directions for Community Colleges, 2006*(135). 5-29

Developmental Working Group (2014). *Developmental education working group report*. Topeka KS: Kansas Board of Regents

Ewell, P., Kelly, P., & Klein-Collins, R. (2008) *Adult learning in focus: National and state-by state-data.* Chicago, Il: Council for Adult and Experiential Learning and National Center for Higher Education Management Systems. Retrieved from http://www.cael.org/adultlearninginfocus.htm

Guison-Dowdy, A., & Patterson, M.B. (2011). *Journeys through college 1: Postsecondary transitions and outcomes of GED® test passers*. Washington, DC: American Council on Education.

Guison-Dowdy, A., & Patterson, M.B. (2011a). *Journeys through college 2: Postsecondary experiences of GED® test passers*. Washington, DC: American Council on Education.

Horn, A.S. (2014). *Determinants of student success: An integrative perspective to orient policy and practice.* Minneapolis, MN: Midwest Higher Education Compact.

IES-NCES (2012). *The condition of education 2012*. Washington D.C.: Department of Education Retrieved from http://nces.ed.gov/pubs2012/2012045.pdf

IES-NCESa (2012a). *Program for international student assessment (PISA) 2012 public-use data* *files*. Washington D.C.: Department of Education. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014028

Kansas Board of Regents. (2010) *Foresight 2020*. Retrieved from http://www.kansasregents.org/

Miller, N. B. (2014). Nontraditional student graduation rate benchmarks. *The Journal of Continuing Education, 62*(3), 141-151.

OVAE-NRS, Kansas Statistical Data, Table 5, 2007, 2008, 2009, 2010. Retrieved from: http://wdcrobcolp01.ed.gov/CFAPPS/OVAE/NRS/tables/index.cfm

Pascarella, E. T. (1985). Students' affective development within the college environment. *The Journal of Higher Education, 56*(6), 640-663.

Patterson, M.B. (2010). *GED® tests passers in postsecondary institutions of up to two years: Following up on enrollment and graduation*. Washington, DC: GED Testing Service.

Patterson, M.B., Zhang, J., Song, W., & Guison-Dowdy, A. (2010). *Crossing the bridge: GED credentials and postsecondary educational outcomes, year one*. Washington, DC: GED Testing Service.

Reder, S. (2007). *Adult education and postsecondary success*. New York: Council for the Advancement of Adult Literacy. http://www.caalusa.org/content/rederpolicybriefrev10807.pdf

Research Allies for Lifelong Learning. (2013). *Crossing the bridge: Adults with GED credentials going to college (2004 – 2010)*. Retrieved from http://www.researchallies.org/Crossing_the_Bridge.html

Song, W., & Patterson, M.B. (2012). *From GED testing to postsecondary education: The role of institutions.* Unpublished manuscript presented at American Educational Research

Association conference, Vancouver, BC.

Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L., & Geman, D. (2005). Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, *21*, 3896–3904.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences, 99*(10), 6567–6572.

U.S. Department of Education, Office of Vocational and Adult Education (2007) *Adult Education and Family Literacy Act of 1998: Annual Report to Congress 2006-2007*. U.S. Department of Education. 2007. Retrieved from www.nrsweb.org/reports/congress_report.aspx.

Wang, H., Dai, Z., Chen, M.S., &Yuan, Z. (2013). TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection. *BMC Medical Genomics*, 6: (Suppl 1): S3.

Zhang, J., Guison-Dowdy, A., Patterson, M.B., & Song, W. (2011). *Crossing the bridge: GED credentials and postsecondary educational outcomes, year two*. Washington, DC: American Council on Education.