Machine Learning for Public Policy

Final project

Group: Save the Students

Member: Xiangyu Ma, Xin Ding

Date: 6/4/2018

Instructor: Rayid Ghani

I. Introduction

For a variety of reasons, we don't like to see people dropping out of high school. It's bad for the individual: the advantages and disadvantages from high-school and college graduation cascade over the lifetime and become enormously consequential on life outcomes (E.g. Ross & Mirowsky 1999; Hout 2012). It's also bad for the state from a resource perspective: as an economist might see it, this is would-be human capital gone a-wasting (Becker 1964). And let's not speak of the issues of inequality that will eventually reverberate from these dropping-outs: the chances are pretty good that these high-school drop outs are not income-blind nor are they race/ethnicity-blind, and in so-doing perpetuate the multiform social injustices that stain our collective conscience (see: the literature on stratification, broadly). It's a situation that calls for effective intervention. But to do so, we need to first be able to identify the vulnerable. Machine learning is a very efficient and scientific way to tackle such problem.

The premise of our project is to use machine learning techniques to produce predictive models that allow for early identification of students at risk of dropping out of high school. Our models use snapshots of students' school record at the end of their 8th grade to generate risk assessments of individual students. The risk assessment comprises (a) a risk score and (b) a risk category assessment. In addition, our model tries to explain how it came to generate any particular risk assessments. These drop-out risk assessments allow educators to focus their interventions on those most in need of them. We introduce a total of three models in this paper. All of them try to maximize recall but make different assumptions about the level of resources available to the school. The first, a KNN model, is best suited to schools with limited resources for intervention. The second, a gradient-boosted tree model, is best suited for schools with moderate resources for intervention. The third, a random forest model, is best suited for schools with liberal resources for interventions.

II. Literature review

Dropping out of school persists as a problem that interferes with educational system's efficiency is the most straightforward and satisfying route to individual educational goals for young people. Studies show that high schools in North Carolina mainly utilize early warning indicators, including attendance rates, behavior incidents, and course performance to identify students with potential adverse school outcomes. With Scantron Analytics tools in hand, schools have the right assessment data for students.

Jan C. Lemon and Joshua C. Watson suggest that school counselors should include interventions targeted on promoting skills which develop purpose in life, compassion for others, moral values, and a sense of oneness with the universe. This would include individual and group counseling with an emphasis on freeing students with "should" and "oughts" in their belief systems and creating plans with new outlooks on how to make the most of the student's academic attributes, personal beliefs, and individual strengths.

Sweeney (2009) stated that there is a close association between wellness and Adlerian concepts. Therefore, individual psychology would be an excellent theory to employ with adolescent clients. Also, it is interesting to note that the significance of these components of wellness indicates that counselors and educators cannot make others do anything that they do not consider personally useful. Thus, the attributes that protect students from making poor academic decisions can only be understood from the aspect of the student's private logic (Sweeney, 2009).

III. Problem formulation & overview of our solution

The early identification of students at risk of dropping out can be framed as a prediction task in machine learning parlance. That is, given a set of features about an individual (which in our case is the snapshot of a student's school record at the end of 8th grade), we are asking our machine classifier to make a judgment if s/he would drop out.

Given our substantive focus, we were primarily interested in recall, the ratio between true positive and the sum of true positives and false negatives: simply put, we wanted to identify as many of the at-risk students to dropping out as possible. There's a simple naive way of maximizing recall: intervene on every single student. But it is clearly infeasible given the resource constraints schools face. What we wanted to do, rather, was to produce a classifier that maximizes recall for a given proportion of population that a school had resources to conduct interventions on.

Following this vein of thought, we constructed models tailored to schools with differing level of resources available for intervention: (a) low resource schools, (b) medium resource schools, and (c) high resource schools. We assume that low resource schools are able to make interventions on 5% of the school population, medium resource schools 15%, and high resource schools 30%.

We ultimately produced a machine learning platform that provides educators with the following:

1. The risk score of a student, scored from 0 to 10.
2. The risk category a student belongs to, whether it be "Low," "Medium," or "High".
3. A brief description of how the two risk assessments above were generated.

IV. Description of our data

This paper uses time-series panel data (2006-2015) from the Muskingum Valley Educational Service Center (MVESC). The MVESC data-set contains the school records of students from 11 high schools in Muskingum Valley, Ohio. We track 5 cohorts of students through their time in high school. We use these snapshot of students' school records, **6332** in total, at the end of their 8th year to build a model that predicts a students' vulnerability to dropping out. We also augment MVESC data with tract-level demographic data from the decennial 2010 U.S. Census.

V. Details of our solution

1) Labels

We labeled students according to the following:

- They were assigned 1 if they dropped out.
- They were assigned 0 if they did not drop out.

In total, we had 6172 observations labelled 0 and 160 observations labelled 1.

2) Features

We had two classes of features, native features and geocoded features. We had 84 features in total.

3) Native Features

Native features refers to features that were derived through trivial to moderate amounts of transformation of the MVESC data. Native features comprise the following:

1. Number of days absent (all).

2. Number of days absent with excuses.

3. Number of days absent without excuses.

4. Number of disciplinary incidents.

5. GPA.

6. Number of classes taken.

7. Number of consecutive lateness to school.

8. 13 variables that indicate disability

9. 3 binary variables that indicate fluency in English.

10. 6 binary variables that indicate ethnicity.

11. 2 binary variables that indicate gender.

12. 11 binary variables that indicate the high school of the students.

13. 1 binary variable that indicators if a student didn't take English, Science, Social Studies, and Mathematics.

14. 12 binary variables for birth month.

4) External features:

We believe demographic information about student’s neighborhood can be very relative and helpful to our study. Thus, we included such geographic information as our external features. Each external features refers to features derived from the spatial data available in MVESC, namely students' home addresses. We did two main things. First, we augmented our MVESC data with 11 tract-level demographic characteristics from the 2010 U.S. Census. They are as follows:

1. % of female high school graduate over 25.
2. % of male high school graduate over 25.
3. % of female with no schooling completed.
4. % of male with no schooling completed.
5. % of female enrolled in school.
6. % of male enrolled in school.
7. Per capita income in the past 12 months.
8. Median household income.
9. Median age.
10. Average household size.
11. Total population.

In order to do so, we first used google API to geocode the addresses for students into coordinates. Then plot such coordinates into points which were later spatial joined with the census tract polygons in Ohio to generate a layer for students in different census tracts. After that, we acquired the census tract and block level demographic data from the 2010 American Community Survey 5-year data and joined them with the layer from the spatial join. Unfortunately, we were not able to find data at census tract level for years before 2009 and therefore settled for using 2010 census data to depict different areas. As a result, we have students divided into census tract with the ten characteristics listed above.

Second, we performed K-Means unsupervised clustering of the home address of the students, creating 300 clusters in total. To avoid generating too many superfluous binary variables, we used a random forest algorithm to select for clusters that had feature importances that were greater than 0.01. Of the 300 cluster IDs we generated, we ultimately retained 18 of them as binary features.

5) Model selection:

Let's start with the process of selecting the best model for high resource schools. First, we split our data into our main training set and test set. The main training set comprises data from 2006-2009, while the main test set contains data from 2010. Second, we tried to identify the best performer for each type of classifier. We tried 6 types of classifiers: (a) Logistic Regression, (b) Support Vector Machine, (c), K-Nearest Neighbors, (d) Decision Tree, (e) Random Forest, and (f) Gradient Boosted Trees. We set up a search-grid of the parameters we want to vary for each classifier.

We then performed cross validation on our main training set with temporal holdouts. This gave us three cross validation sets:

1. Train on 2006; test on 2007.

2. Train on 2006-2007; test on 2008.

3. Train on 2006-2008, test on 2009.

For each cross validation set, we calculated the recall when up to 30% of the population is intervened upon. Then, we proceeded to identify the best model for each classifier. The best model was defined as the model with the highest mean recall across the 3 cross validation sets when up to 30% of the population is intervened on. This gave us 6 models, the best performers from each type.

Next, we tested these 6 models against the main test set. The best model for high resource schools was the model with the highest recall when up to 30% of the test set population was intervened on.

We repeated the process to find the best model for medium resource and low resource schools (defined as being able to intervene on up to 15% and 5% or the student population, respectively). The only difference was our evaluation metric for best classifier. In the case of the medium resource school, its highest recall was at 15% of population; in the case of the low resource school, its highest recall was at 5% of population.

VI: Discussion

1) Overview of the models:

We produced 3 models, each tailored to schools with differing amounts of resources. Even though three models were based on different types of classifiers, they ultimately take the same inputs and produce the same outputs. To use our models, an educator feeds in a student's student lookup ID, and our model would output 3 things: (1) the student's risk score, (2) the student's risk category, and (3) how our classifier makes the decision.

The risk score was generated by scaling our model's probability prediction to a 0-10 scale. The risk category came from our attempt to put these risk scores into meaningful buckets. For now, scores of 9.5 and above are considered "high risk," scores of 8.0-9.5 are considered "medium risk," and scores of 7.9 and below and considered low risk. We recognized that these risk buckets may seem arbitrary at present. In practice, we'd work with schools to find what kinds of risk buckets make the most sense to them. Finally, we believe that it's extremely important to explain how our classifier makes our decisions. It greatly helps interpretability, and it can help engender greater trust in our models (Ribiero et al. 2016). Our high resource and medium resource models, since they are both tree-based models, try to explain their decisions by making reference to the most important features. Our low resource model is based on a KNN model — what we try to do here, then, is to show the educator a small sample of the nearest neighbors our model relied on to make its risk assessment for any particular student.

2) The model for high resource schools:

This is based on a Random Forest Classifier with the following parameters:

min_samples_split = 2,

max_features = "sqrt",

n_estimators = 1000,

max_depth = 5, criterion = "entropy",

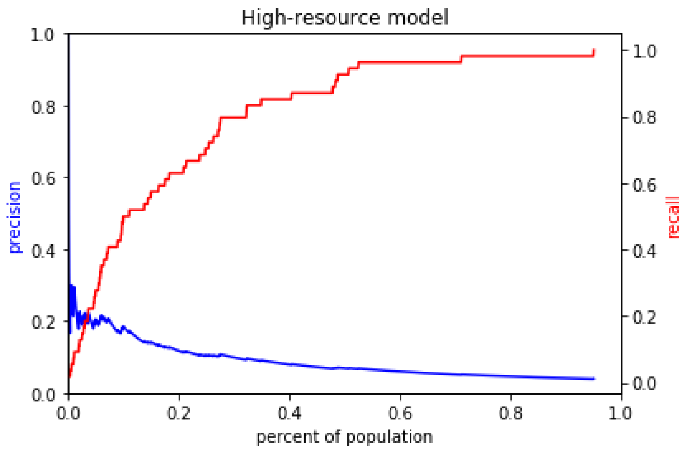
n_jobs = -1,

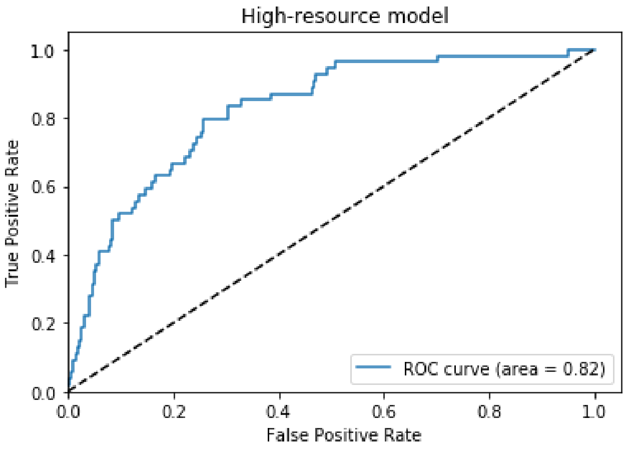
random_state = 614).

Model performance:

When targeting 30% of the population,

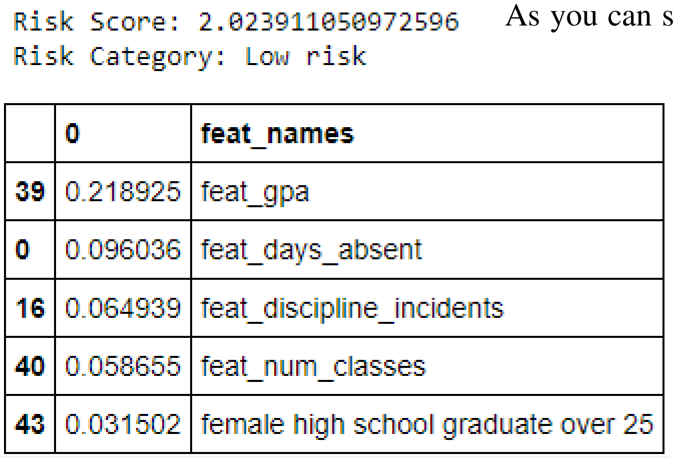
it has a recall of 0.80, precision of 0.10, and a ROC-AUC score of 0.76.





The high-resource model in action:

Let's pretend the first student, student X, from the main test set is a student from out of sample that an educator wants a prediction for. When the educator feeds the data this student X's record. This is what he'd get back as output.



```
Risk Score: 2.023911050972596
Risk Category: Low risk
```

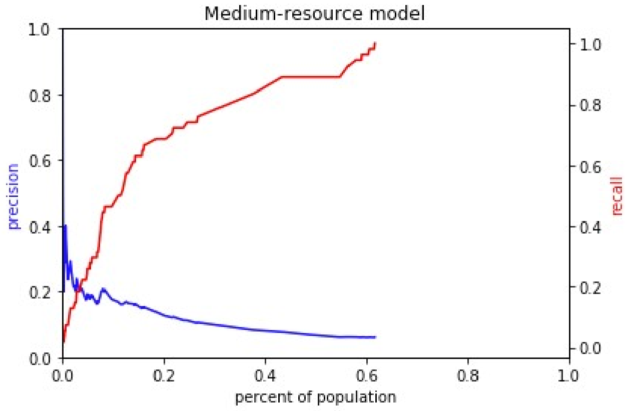
	0	feat_names
39	0.218925	feat_gpa
0	0.096036	feat_days_absent
16	0.064939	feat_discipline_incidents
40	0.058655	feat_num_classes
43	0.031502	female high school graduate over 25

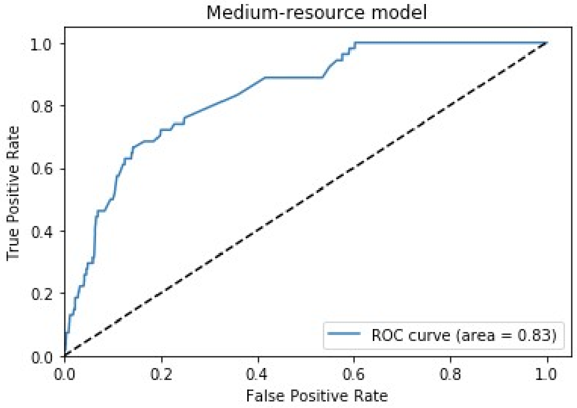
As you can see here in this case, our model predicts that student X is at low risk of dropping out, with a risk score of 2.02 out of 10. In addition, it tells the educator that the decision was based primarily on GPA, number of days absent, number of disciplinary incidents, number of classes taken, and the % of female high school graduate over 25 in the census tract student X lives in.

3) The model for medium resource schools:

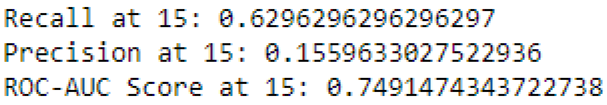
Our model for medium resource schools is based on a Gradient-Boosted Tree with the following parameters: loss = 'exponential', learning_rate = 0.01, n_estimators = 1000, max_depth = 1 and random_state = 614.

Model performance:





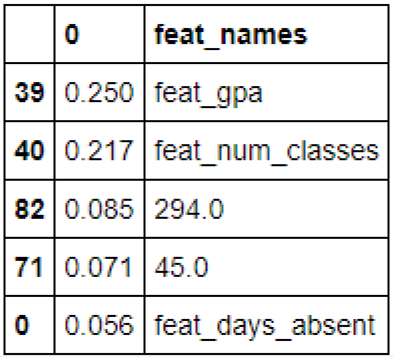
When targeting 15% of the population, it has a recall of 0.63, precision of 0.16, and a ROC-AUC score of 0.75.



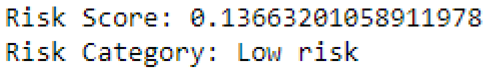
```
Recall at 15: 0.6296296296296297
Precision at 15: 0.1559633027522936
ROC-AUC Score at 15: 0.7491474343722738
```

Example of medium-resource model in action:

In this case, our model predicts that student X is at low risk of dropping out, with a risk score of 0.14 out of 10. In addition, it tells the educator that the decision was based primarily on GPA, number of classes taken, whether or not s/he lives in cluster IDs 294 and 45, and the number of days absent.



	0	feat_names
39	0.250	feat_gpa
40	0.217	feat_num_classes
82	0.085	294.0
71	0.071	45.0
0	0.056	feat_days_absent

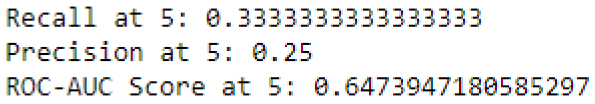


```
Risk Score: 0.13663201058911978
Risk Category: Low risk
```

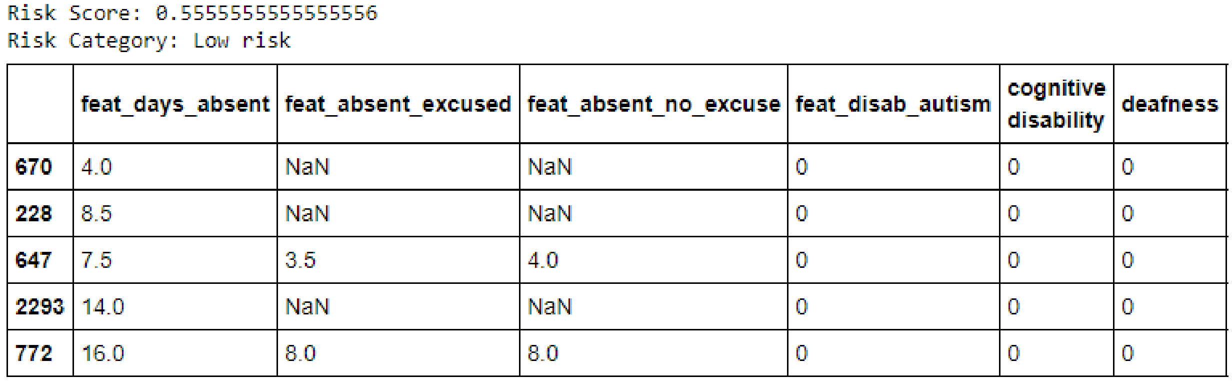
4) The model for low resource schools:

Our model for low resource schools is based on a K-Nearest Neighbors model with the following parameters: n_neighbors = 50, metric = "euclidean", weights = "distance", n_jobs = -1.

Model performance:



```
Recall at 5: 0.3333333333333333
Precision at 5: 0.25
ROC-AUC Score at 5: 0.6473947180585297
```

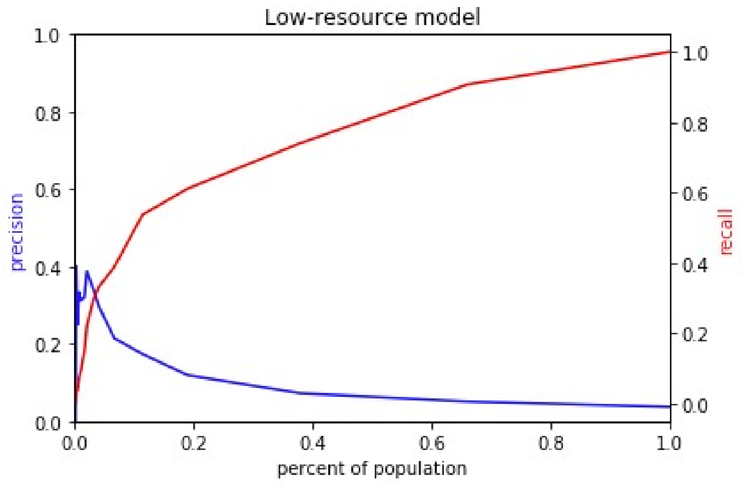


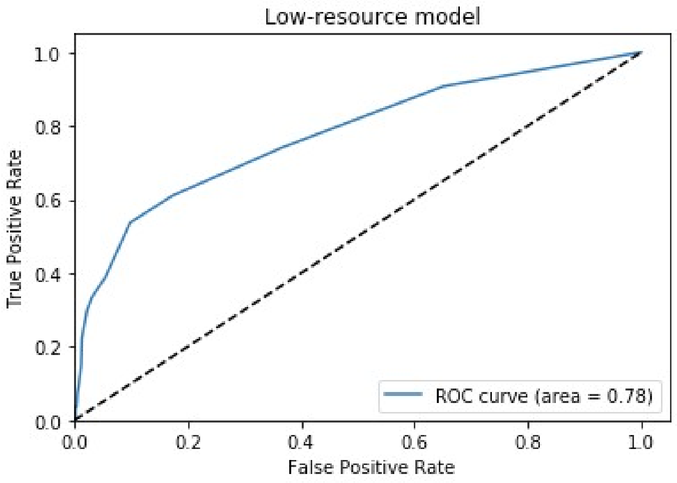
```
Risk Score: 0.5555555555555556
Risk Category: Low risk
```

	feat_days_absent	feat_absent_excused	feat_absent_no_excuse	feat_disab_autism	cognitive disability	deafness
670	4.0	NaN	NaN	0	0	0
228	8.5	NaN	NaN	0	0	0
647	7.5	3.5	4.0	0	0	0
2293	14.0	NaN	NaN	0	0	0
772	16.0	8.0	8.0	0	0	0

When targeting 30% of the population, it has a recall of 0.33, precision of 0.25, and a ROC-AUC score of 0.65.

The low-resource model in action:





(other columns not shown here for limited space)

In this case, our model predicts that student X is at low risk of dropping out, with a risk score of 0.56 out of 10. In addition, it shows the educator a subsample (5) of the 100 nearest neighbors it used to make its prediction.

VII: Policy recommendations

Our goal was to help reduce high school dropout rates. To do so, we developed three types of early identification systems, one for low-resource schools, one for medium-resource schools, and one for high-resource schools. These early identifications use the 8th grade school records of high schoolers to predict their likelihood of dropping out of high school. Educators using our models will receive a risk assessment coupled with an explanation of how it was made.

Of course, simply producing a predictive model alone can't reduce high school dropout rates. For our goal to be achieved, we must (a) promulgate its adoption and (b) couple it with an effective intervention program.

Our models have relatively high recall considering the fact that the dropouts are usually the rare cases. We would suggest to use this model as a prediction/reference for intervention programs, but not as evidence for dropout prediction.

It is also worth noticing that the features highlighted by our model should not be picked out individually for any causal inference.

VIII: Limitations

Limited Data sources: 6,332 data entries; missing values. ACS data: ACS data only goes down to census tract level, which may not precisely reflect each individual's circumstances; census data in 2010; For unsupervised clustering, we need to further validations; Try to enhance our precision in the future; we also need access to current heuristics for more meaningful comparisons of lift;

We also want to be cautious and not oversell our results too much. We think our results seem good — but the proof is really in its lift relative to the extant heuristics schools use. How much better (or worse) are our models compared to a simple rule-based model schools use? Without that information at hand, our talk may seem like mere bluster.

We also assume that there are existing interventions that are effective. But are there? And more importantly, would these interventions be effective on the at-risk students our model identifies? Identifying the vulnerable might not be very useful if it turns out that we can't do very much at all. Our literature review into this has been ankle-deep at best. We think the answer is yes — there has to be, right? — but we aren't a 100% sure.

IX: References

Aguiar, et al. Who, When, and Why: A machine learning approach to prioritizing students at risk of not graduating high school on time. 2015

Becker, Gary S. 1994. Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. *University of Chicago Press.*

City Year, Scaling City Year's Impact: Growth Plans to Reach 50% of Off-Track Students in City Year's 20 U.S. Locations(2015)

Hout, Michael. 2012. "Social and Economic Returns to College Education in the United States." Annual Review of Sociology*38(1):379–400.*

REL Southwest , Applying an On Track Indicator for High School Graduation: Adapting the Consortium on Chicago School Research Indicator for Five Texas Districts (2011).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." ArXiv*:1602.04938 [Cs, Stat].*

Ross, Catherine E. and John Mirowsky. 1999. "Refining the Association between Education and Health:

The Center for Social Organization of Schools, Johns Hopkins University. Advancing the "Colorado Graduates" Agenda: Understanding the Dropout Problem and Mobilizing to Meet the Graduation Challenge (2009)

The Effects of Quantity, Credential, and Selectivity." Demography *36(4):445–60.*

The Rennie Center for Education Research & Policy, Meeting the Challenge: Fiscal Implications of Dropout Prevention in Massachusetts (2011)