Beyond Early Warning Indicators: High School Dropout and Machine Learning♠

Dario Sansone[1]

October 2017

Abstract

This paper provides an algorithm to predict which students are going to drop out of high schools relying only on information from 9th grade. It verifies that using a parsimonious early warning system - as implemented in many schools - leads to poor results. It shows that schools can obtain more precise predictions by exploiting the available high-dimensional data jointly with machine learning tools such as Support Vector Machine, Boosted Regression and Post-LASSO. It carefully selects goodness-of-fit criteria based on the context and the underlying theoretical framework: model parameters are calibrated by taking into account policy goals and budget constraints. Finally, it uses unsupervised machine learning to divide students at risk of dropping out into different clusters.

Keywords: high school dropout; machine learning; big data

JEL: C53; C55; I20

♠ I am grateful to Garance Genicot, Francis Vella, Laurent Bouton, Daniel Ackerberg, Pooya Almasi, Mary Ann Bronson, Nick Buchholz, Benjamin Connault, Francis DiTraglia, Luca Flabbi, Myrto Kalouptsidi, Ivana Komunjer, Gizem Kosar, Arik Levinson, Whitney Newey, Elena Arias Ortiz, Franco Peracchi, Mariacristina Rossi, John Rust, Bernard Salanié, Shuyang Sheng, Arthur van Soest, Allison Stashko, Basit Zafar and participants to the 2017 Stata Conference, the 2017 GCER Alumni Conference, the George Washington University SAGE Conference, the 2017 APPAM DC Regional Student Conference, and the Georgetown University EGSO seminar for their helpful comments. I am also grateful to John Rust and Judith House for their technical support. The usual caveats apply.

[1] Corresponding author. Georgetown University (Department of Economics, ICC 580, 37th and O Streets, NW, Washington DC 20057-1036, USA). E-mail: ds1289@georgetown.edu

1. Introduction

High school dropout is a key issue in the US educational system and it has been extensively analyzed by researchers (De Witte et al., 2013; Murnane, 2013). Indeed, the percentage of students who graduated with a regular high school diploma within four years of starting 9th grade was only 83.2% in the academic year 2014/15. These figures are in line with the upper-secondary graduation rate of 82% (79% for male, 85% for female) reported by the OECD (2016), below the average of 85% in 2014 among advanced economies, and far from the achievements of Germany (91%), Japan (97%) and Finland (97%). Furthermore, there are substantial racial and geographical gaps within the US (IES, 2016).[2] Failing to graduate from high school can also have deep individual costs. Indeed, by 2020 only 12% of all jobs in the economy will require less than a high school education (Carnevale et al., 2013). Schooling also offers several nonpecuniary benefits ranging from health to happiness, marriage, trust, and work enjoyment (Oreopoulos, 2007; Oreopoulos and Salvanes, 2011).

This paper shows how machine learning (ML) can be applied in education. In particular, the goals are to create a model which allows schools to identify students at risk of dropping out using only information from their first year of high school, as well as to show how ML can be used to identify top predictors and heterogeneities among students. It demonstrates that trying to predict vulnerable students using traditional predictors leads to detect only a small fraction of the students who actually end up dropping out of high school. Despite this disappointing result, schools often rely on these few early warning indicators to identify students who are struggling academically (O'Cummings and Therriault, 2015). Indeed, educators are advised to focus only on attendance, school behavior and course grades to find weak students, even if – according to What Works Clearinghouse (Rumberger et al., 2017) - there is minimal empirical evidence to support this recommendation. In contrast with these practices, this paper explains how schools can exploit big data, jointly with ML techniques, to substantially improve these predictions and thus target weak students.

After having identified the students who are at risk of dropping out, the paper demonstrates how to use unsupervised ML to cluster such individuals into different groups. This division has two advantages. First, it stresses that these students are not a homogeneous group, therefore policy-makers need to design specific interventions to efficiently target them. In other words, the ML algorithm may classify some students as at-risk because they are academically weak, while others may be predicted as dropouts because they live in unsafe neighborhood or they come from very poor households. The latter group would require different programs than the first one. ML can therefore be used to identify students at-risk, as well as to help design treatments appropriate for each sub-population. Second, it is possible to evaluate how a policy affects differently the students in the various clusters. Indeed, any treatment may have different impact depending on student gender, race, ability, income, as well as by sub-populations. In this way, it is possible to

[2] It should be stressed that graduation rates, racial differences and time trends are extremely sensitive to the sample used, as well as to whether GED recipients are counted as high school graduates (Heckman and LaFontaine, 2010).

estimate heterogeneous effects not only on different demographic groups, but also on multidimensional groups.

This paper is related to the emerging literature in ML. The main focus of econometric techniques is causal inference, i.e. to provide unbiased or consistent estimates of the impact of a variable x on an outcome y. On the other hand, ML is more appropriate for prediction purposes since its goal is to maximize out-of-sample prediction. Algorithms can indeed identify patterns too subtle to be detected by human observations (Luca et al., 2016), thus outperforming econometric models built using heuristic or theory-based approaches. While ML applications are widespread in computer science, its use in economics has been quite limited so far, although there are several policy-relevant issues that do not require causal inference, but rather accurate predictions (Kleinberg et al., 2015). In fact, ML is gaining momentum in this field (Belloni et al., 2014; Varian, 2014; Mullainathan and Spiess, 2017) and scholars have started to use these algorithms in education for teacher tenure decisions (Chalfin et al., 2016), as well as to reduce dropout rates in college (Aulck et al., 2016; Ekowo and Palmer, 2016).

One disadvantage of mechanically using ML techniques off-the-shelf to tackle classification problems, i.e. applications where the dependent variable is discrete, is that there is no unique method to measure performances. Indeed, practitioners generally adopt rule-of-thumbs and criteria such as pseudo-R^2 and accuracy (Bowers et al., 2013), but they do not actually justify the reason behind such choice. On the other hand, in line with Subrahmanian and Kumar (2017), this paper build an economic model in order to derive a criterion consistent with the school objective function which can be used to compare the performances of different algorithms. In other words, the goals of the school and its budget constraint are taken into account while calibrating the algorithm to maximize performances. Therefore, this analysis provides a microeconomic foundation to the choice of the particular criterion used in the paper to evaluate the algorithms.

Despite this limitation, the ML approach provides several advantages. First, it offers an inexpensive alternative to the numerous tests and assessments that are used since kindergarten to sort and categorize students (Shields et al., 2016). Second, since our algorithm uses only information from 9^{th} grade, school counselors and teachers may detect in advance weak students before it is too late to intervene.

Finally, this paper is related to previous studies – mainly written by non-economics in technical reports – on predicting high school dropout (Rumberger and Lim, 2008; Bowers et al., 2013; Adelman et al., 2017). There have also been some preliminary attempts by data analysts to predict high school dropouts using ML algorithms. Sara et al. (2015) trained ML algorithms using few variables from administrative data in Denmark to predict dropout three months later. Aguiar et al. (2015) introduced ML to predict dropout in a U.S. school district using few early warning indicators and demographic variables, while Knowles (2015) used ML to improve the dropout early warning system in Wisconsin.

As already mentioned, this paper expands this literature in several ways. First, it introduces a theoretical model to justify the goodness-of-fit criterion used to evaluate different specifications. Second, it strongly warns against the risks of using few early warning indicators and it relies instead on a large set of variables. Third, it investigates the performances of alternative ML algorithms and uses them to predict dropout years – not months – later. Fourth, it applies unsupervised ML for the first time in the educational context. Last but not least, it is the first one to use a recent U.S. nationally representative data set, thus reducing the external validity concerns raised for local analysis.

2. Data

2.1 Data Source

The High School Longitudinal Study of 2009 (HSLS:09) is a panel micro study including around 26,000 students in 9th grade from about 940 participating schools in 2009. The survey design has two levels: first, schools were selected at the national level (both private and public). Second, around 30 students in each school were randomly selected among 9th graders. Among eligible students, around 21,440 students responded.[3]

In the first round, information was collected from the selected 9th graders, their parents, math and science teachers, school administrators and lead school counselors. The parent questionnaire was completed by the parent or guardian most familiar with the 9th grader's school situation and experience. The students were interviewed between September 2009 and April 2010. The first follow-up was in the spring of 2012, while a brief update was conducted in 2013 (summer and fall) to record students' postsecondary plans. In 2012 students, parent, school administrators and counselors were interviewed again, but this wave did not include new questionnaires for teachers. Finally, in 2013 only students and parents were interviewed.

A math assessment was administered to the students in 9th grade (2009) and in 11th grade (2012). Data are also available from the student transcripts including their GPA, their AP class grades, their SAT scores, and the number of credits taken in each subject during high school.[4]

From a policy perspective, the use of the HSLS:09 implies another substantial contribution of this paper: the results presented in this paper do not only focus on the general issue of high school dropout, but are derived from data regarding a recent cohort, thus offering a new perspective on Millennials and their educational choices. Indeed, most of the previous literature has exploited data such as the NLSY79, which are attractive since they contain a rich variety of information and span over several decades, but they may estimate parameters which have changed with the new generations, thus lacking external validity.

[3] Around 550 students were not deemed capable of completing the questionnaire because of limited English proficiency or disability (mental, emotional or physical).

[4] Additional documentation about the HSLS:09 can be found in technical reports provided by the US Department of Education (Ingels et al., 2011; Ingels et al., 2014; Ingels et al., 2015). For security reason, all sample size numbers have been rounded to the nearest 10.

2.2 Outcome Variable and Predictors

The aim of the first section of the paper (Section 3) is to predict who is eventually going to drop out of high school using only information available in 9th grade, i.e. in the first year of high school. To begin with, it should be noted that 45% of the schools in the sample had a formal dropout prevention program in 2009. These programs included a variety of initiatives: the most common were tutoring and graduation counseling, but some schools also offered job counseling, childcare for children of students, occupational focused courses, or even incentives for better attendances and classroom performance. When school counselors were asked how students were selected in order to participate in these programs, the two most common answers indicate a focus on individuals with poor grades (93%) and low number of credits (89%).

The main outcome variable used in the empirical analysis is *Ever dropout*. This is an indicator variable equal to one if there is at least one know dropout episode regarding the student, zero otherwise. It is important to stress that alternative completers (such as GED recipients) are considered as dropouts, which is in line with the literature stressing the differences between GED recipients and high school graduates (Heckman and Rubinstein, 2001; Heckman et al., 2011; Zajacova, 2012). Even more importantly, non-respondents are counted as zero.[5]

Almost 11% among the interviewed students had at least one known dropout episode before the second follow-up interview. It is also important to note that the vast majority of the schools in the sample (82%) had at least one student with a recorded dropout episode. Therefore, in line with the findings from other studies (Adelman et al., 2017), dropouts are not concentrated only in few schools, thus merely targeting low-performing schools would lead to substantial misallocation of resources.

3. Predictions

3.1 Technical Considerations

Before showing the results from the prediction analysis, it is important to highlight few technical points. The first one is how to avoid over-fitting, i.e. to have high in-sample predictive power, but low out-of-sample one. For instance, if the true relation between y and x is quadratic, a linear model would be an under-fit (high bias), while estimating a 4th degree polynomial would lead to an over-fit (high variance). As suggested by (Ng, 2016), the solution is provided by dividing the data into three samples. The training sample (60% of the data) is used to estimate the algorithm. The optimal model parameters (such as the penalization term in LASSO) are selected using a grid-search in order to maximize the performances in the CV sample (20% of the data, around 4,290 observations). Therefore, the risk of overfitting is reduced by estimating the model using the training data and measuring the performances using the CV sample. Finally, the out-of-sample performances are reported using the test sample (20% of the data). This last – less

[5] The Online Appendix discusses additional robustness checks where non-respondents are counted as missing.

common - step is required since an extensive grid-search may still lead to overfitting the CV sample.

The main concerns with this simple form of CV are that not all data are exploited to calibrate the model and, in case of relatively small samples as in this case, there is a risk that outliers may be overrepresented in one of the three samples. These issues can be avoided using 5-fold CV. In fact, the data have been divided into five sets and combined in all possible ways in order to create five different splits among train, CV, and test samples. We have then estimated in-sample and out-of-sample performances five times, one for each data split, and reported the 5-fold average out-of-sample performances.

The second technical point worth mentioning is that there is not a unique measure of performances when the dependent variable y is binary. Indeed, while the Mean Square Error (MSE) or the R^2 offer a clear metric when the dependent variable is continuous, such criteria are not appropriate in classification problems. There are two classes of indices in this setting. The first one, which includes the pseudo-R^2 and the McFadden-R^2, compares the performances of the algorithm with the prediction of a simple model which contains only a constant. The second class comprises all the indices which compare observed values with predicted ones. The usual starting point in this case is the so-called "confusion matrix", which tabulates the frequencies of the actual values of the dependent variable against the values predicted by the model.

		Predicted values	
		0	1
Actual values	0	Correct$_0$ (c_0)	Wrong$_1$ (wr_1)
	1	Wrong$_0$ (wr_0)	Correct$_1$ (c_1)

Which can be interpreted as:

		Predicted values	
		0	1
Actual values	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

The most frequent criterion used to evaluate a classification algorithm is the accuracy rate:

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ observations\ (n)} = \frac{c_1 + c_0}{n}$$

However, when classes are imbalanced as in our case, i.e. when the number of positive values (n_1) of the dependent variable - i.e. the number of high school dropouts - is much smaller than the number of zeros (n_0), such criterion is not appropriate since a naïve model with just a

constant would reach a very high accuracy rate. In these cases, the indices which are commonly used are:

$$Precision\ (or\ Positive\ Predicted\ Value) = \frac{TP}{TP + FP} = \frac{c_1}{c_1 + wr_1}$$

$$Specificity = \frac{TN}{TN + FP} = \frac{c_0}{c_0 + wr_1}$$

$$\varphi = \ Recall\ (or\ Sensitivity) = \frac{TP}{TP + FN} = \frac{c_1}{c_1 + wr_0}$$

Other available criteria are the F_1-score and the Negative Predicted Value. Given this variety of measurements, most analysts tend to arbitrary pick one or two of them following common practices and rules of thumb. In what follows, we focus on the recall rate since we believe predicting that a student is not at risk when he or she actually ends up dropping out is an error which can have bigger consequences that the opposite mistake, i.e. when a student who graduates from high school is identified as at-risk. In Section 4, we justify this choice using a microeconomic model and we include a budget constraint in the analysis.

Finally, almost all algorithms (with the notable exception of SVM) produce by default predicted probabilities. We initially follow the convention to predict one when such probability is equal or greater than 0.5, zero otherwise. This is in line with the Bayes classifier (Hastie et al., 2009): the accuracy rate is maximized by assigning each observation to the most likely class, given its predicted probabilities. Lower thresholds lead to higher recall rates, but lower accuracy. Subsequently, we choose such cut-off during the CV procedure in order to optimize the school objective function. We show in Section 4 how this procedure is related to the ROC curve, which can be generated non-parametrically using each possible predicted probability as a classification threshold and computing the corresponding sensitivity and 1-specificity, thus highlighting the trade-off between these two criteria. The area under such curve (AUC) is also commonly used as a performance criterion.

3.2 A Basic Model

As discussed in Section 2.2, most of the schools select students who need to participate in dropout prevention programs based on student past performances (GPA and number of credits). Therefore, a natural way to start the analysis is to test the power of these predictors. In other words, we have estimated a simple logit model using as regressors student past performances, school attendance and behavior, as well as all the others variables highlighted in the literature: demographics, school characteristics, and family background.[6] As shown in Table 1 Model 1, the

[6] In particular, we have selected the following 28 variables: student gender, race, language, school region, urbanicity, school climate, household income, number of household members, no mother/father in the household, mother/father high school dropout, mother/father employed, student has repeated a grade, 9th grade math test score, 9th grade GPA, 9th grade number of credits, school attendance, school suspension. The Online Appendix includes a detailed description of all the variables used in this section. In order to compare results with the ML algorithms, we have used also in this case a 5-fold CV procedure. Moreover, in order to maintain the same number of observations across specifications, we have imputed missing values to zero

performances are strikingly low: even if the average out-of-sample accuracy rate is almost 90%, the recall rate is only 15%, meaning that only a small percentage of the students in the test sample who do eventually drop out are identified as at-risk. These performances are even worse for the OLS and Probit estimates (Model 2 and 3 respectively). Moreover, this does not depend on the sample size: similar accuracy and recall rates are obtained also using only random subsets of the train sample (e.g. 30%, 50%, 80%). Even if we estimate the Logit model using all the observations, the in-sample recall rate is only 17% (92% accuracy, 0.81 AUC). This implies that collecting data on additional students would not improve our predictions: the algorithm is suffering from high bias (under-fitting), thus including more training observations would not solve this issue.

We can also add key interactions terms to take heterogeneity into account and lead to a more flexible functional form. For instance, boys and girls may have different likelihood of dropping out based on their ethnicity, household composition or parental employment. Nevertheless, as shown in Model 4, the addition of 14 interaction terms does not improve performances.

Table 1: Basic Model - 5-Fold Average

		Inputs			Performances		
	Algorithm	Individual	School	Interactions	AUC	Accuracy	Recall
1	Logit	✓	✓		0.80	89.9%	15.2%
2	OLS	✓	✓		0.79	89.6%	6.4%
3	Probit	✓	✓		0.80	89.9%	13.9%
4	Logit	✓	✓	✓	0.80	89.9%	15.5%

Note: *Individual* indicates that the algorithm has used as inputs the selected variables from the student and parent questionnaires. *School* refers to selected inputs from principal, while *Interaction* indicates that the algorithm includes two-way interaction terms between gender, race, income, GPA and family characteristics.

We can conclude from this section that schools cannot use basic statistical techniques and rely only on traditional demographic characteristics, previous student achievements, school attendance and behavior in order to identify weak students. In other words, while it is true that graduation rates are lower, for instance, among black students or children in poor single-parent households, these variables are not enough to capture the variety of circumstances which lead students to halt their education career. Similar poor results have also been found in other studies on early warning indicators actually adopted in school districts (Deussen et al., 2017). The next sections show how schools can use the combination of high-dimensional data and ML algorithms in order to improve their predictions.

3.3 Machine Learning: Brief Introduction

This section briefly describes the ML algorithms employed in the paper. A more detailed technical explanation is provided by Hastie et al. (2009), as well as by Ng (2016). The Online Appendix includes detailed technical implementation information.

---

and added an indicator variable for the missing items. Performances for the models without imputations are comparable to those in Table 1: the k-fold average accuracy for the Logit model is 91.6%, while recall is 17.2% and AUC is 0.80.

Machine Learning is the science of getting computers to learn without being explicitly programmed. Standard econometric techniques, i.e. regressions, are considered supervised algorithms. In other words, supervised algorithms are provided with a certain number of "right" answers, i.e. actual y associated with certain x, and are asked to produce other correct answers, i.e. to predict new y given other combinations of x. On the other hand, unsupervised learning algorithms derive a structure for the data without necessarily knowing the effect of the variables. Supervised ML are applied in Section 3.4 to predict high school dropouts, while unsupervised ML are used in Section 5 in order to divide the students predicted to be at risk of dropping out into different groups.

When considering all the relevant variables collected during the baseline interview and all the possible answers, the number of predictors is more than 1,700.[7] Consequently, once we start to also include some interaction terms between the most important predictors, the number of variables can easily reach several thousands. Therefore, given the limited number of observations, it is not possible to include all of them in an OLS or Logit model. Indeed, adding too many variables to these models would lead to over-fitting. Furthermore, OLS cannot be used when the number of regressors is higher than the number of observations. ML algorithms are the appropriate tools to deal with these high-dimensional data sets.

LASSO is an example of a model selection algorithm: it identifies the variables with the highest predictive power, while constraining all the other coefficients to zero. It can be obtained by adding a penalization term λ to the OLS objective function[8]:

$$\hat{\beta}(\lambda) = \underset{\beta \epsilon \mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \|\beta\|_1$$

$$\|\beta\|_1 = \sum_{j=1}^{k} |\beta_j|$$

Since it introduces bias in the coefficients, it is advisable to run a Post-LASSO OLS regression using only the variables selected by the ML algorithm. LASSO is one of the most common ML techniques. Indeed, it is one of the first tools taught in ML courses (Hastie et al., 2009), and it has also been used by economist for selecting the appropriate set of controls when estimating causal effects (Belloni et al., 2014). However, the key assumption is that the data generating process is sparse: only a small subset of variables is assumed to have high predictive power. This may not be realistic in many economic applications (Giannone et al., 2017).

[7] The Online Appendix includes a detailed list of all the variables used as inputs in the ML algorithms. These include, among the others, student demographics, past performances, future expectations, behavior, school identity, relationships with adults and peers, opinions about 9th grade teachers, household composition, mother/father education and working history, household welfare, school characteristics, and information about teacher and student body.

[8] The usual caveat in these techniques is to normalize with zero mean and unit variance all the variables, or to restrict their domain between zero and one, so that the regularization is not inflated by the different scale of the variables. Both methods should work correctly (Guenther and Schonlau, 2016).

Support Vector Machines (SVM) can be seen as a modified Penalized Logistic Regression with the addition of kernels in the objective function:

$$\hat{\beta}(C) = \underset{\beta \epsilon \mathbb{R}^k}{\operatorname{argmin}} C_1 \left[\sum_{i=1}^{n} y_i \, max\{0, 1 - K_i'\beta\} + (1 - y_i) \, max\{0, K_i'\beta - 1\} \right] + \|\beta\|_2$$

Where C_1 is the penalization parameter. Kernel functions allow SVM to be extremely flexible, but at the cost of interpretability. The most common kernel is the Gaussian one, although we have also considered the sigmoid kernel.

$$K_{Gaussian}(x_1, x_2) = similarity(x_1, x_2) = exp\left(-\frac{\|x_1 - x_2\|}{2\sigma^2}\right)$$

$$K_{sigmoid}(x_1, x_2) = \tanh(\theta + \gamma x_1' x_2)$$

It can be shown mathematically that the SVM is a Large Margin Classifier. In other words, SVM selects the curve (or hyperplane) which separates the two classes with the maximum margin. Researchers have shown that SVM can achieve higher performances than other ML algorithms (Maroco et al., 2011).

Boosting (also called Boosted Regression) can be seen as a combination of a sequence of classifiers where, at each iteration, observations which were misclassified by the previous classifier are given larger weights. Indeed, a simple version of boosting can be illustrated by the AdaBoost algorithm:

1. All observations have initially the same weight $1/n$
2. Estimate the first classifier (e.g. a logistic regression or tree) with the equally weighted data
3. Compute the classification errors, increase the weights of the misclassified observations
4. Estimate the second classifier with the new observation weights
5. Repeat steps 3-4 until you have M classifiers
6. Combine all the M individual classifiers by giving more weight to the classifiers with better predictions.

In other words, this algorithm learns from past mistakes and updates its predictions over time. The underlying idea is that combining simple algorithms such as regression trees can lead to higher performances than a single, more complex, algorithm such as Logit. A regression tree optimally partitions the covariate space into a set of rectangles and it then fits a simple model (constant) to each rectangle. Therefore, the estimated function is just the average of the outcomes included in a particular rectangle. In other words, the partition can be described as a series of if-then statements, and it can be visualized by a graph that looks like a tree. The simplest possible tree is called tree stump and it contains only one split and two terminal nodes. Tree stumps tend to work well in boosting (Schonlau, 2005).

Boosted regression is actually implemented using the algorithm introduced by Friedman et al. (2000) since they were able to reinterpret it in a likelihood framework, thus making it comparable to the objective function of the OLS or Logit model. Boosting have also been found to have superior performances than other ML algorithms in many simulations (Bauer et al., 1999; Bauer et al., 1999) and has already been used by economists in other applications (Chalfin et al., 2016).

One of the reasons we present results for these different algorithms is that they offer different combination of interpretability and flexibility. Indeed, post-LASSO is easily interpretable since it just selects a subset of variables to use them as predictors in an OLS model, making simple to understand the contribution of each variable. On the other hand, SVM and Boosting are among the most flexible algorithms: they are able to fit an extremely large variety of functional forms, but they are "black boxes" which do not provide detailed information on how the inputs have been combined, thus lacking transparency.

As also stressed in Aguiar et al. (2015), previous studies predicted high school dropouts by combining early warning indicators. However, these studies had to decide whether to predict that a student would dropout based on the *intersection* of two or more indicators (e.g. low grades and low school attendance), or based on the *union* of these indicators (e.g. low grades and/or low school attendance). The advantage of ML is that researchers do not have to specify a priory how the variables interact among themselves: the algorithm selects the optimal combination with the highest predictive power.

Other studies have use principal component analysis as a preliminary step to combine several variables into a few indicators and to use them as predictors in a Logit model (Adelman et al., 2017). However, this technique provides a dimensionality reduction by only summarizing the joint distribution of a set of variable. There is no guarantee that such transformation preserves the signal which has the most predictive power since this is not the objective of the technique. In other words, the dimension captured by a principal component may not be the most relevant one when predicting dropout (see also Witten and Tibshirani, 2010). On the other hand, ML algorithms can handle high-dimensional data, thus there is no need to reduce the number of predictors ex-ante, and it is possible to fully capture the predictive power of each variable.

3.4 Machine Learning: Results

Table 2 reports the 5-fold out-of-sample performances of all the ML algorithms. All relevant predictors from 9th grade have been included as inputs in Models 1-5. Since the objective is to reduce dropout subject to the limited resources available, the algorithms has been calibrated in order to maximize the recall rate in the CV sample subject to a minimum accuracy rate (0.89, thus similar to the accuracy of the basic models in Table 1). In other words, the parameters in the ML algorithms has been chosen to identify as many dropouts as possible while keeping the number of false positive as low as possible. Section 4 formally justifies this calibration procedure.

As already mentioned in Section 3.3, LASSO tackles high-dimensional data by selecting the most important predictors among all the inputs. These variables are then used as regressors in an OLS (Model 3) or Logit (Model 4) specification. As reported in Table 2, Post-LASSO algorithms manage to substantially increase the recall rate up to 23% - almost 8 percentage points increase compared to the basic model, i.e. a 51% improvement – while maintaining a comparable accuracy rate. It is remarkable that, even if these performances are far from perfect prediction[9], these improvements can be obtained by schools districts with rich data set at no extra cost by just including additional variables in their models.

Similar performances are obtained by SVM (Model 1), Boosting (Model 2) or by including interaction terms in the Logit Post-LASSO algorithm (Model 5). Including school fixed effects (FE) in a Logit model together with the individual variables selected by LASSO produces higher recall rate, but at the cost of lower accuracy (Model 6).[10]

Table 2: ML - 5-Fold Average

		Inputs				Performances		
	Algorithm	Individual	School	Interactions	School FE	AUC	Accuracy	Recall
1	SVM	✓	✓			0.77	89.1%	21.7%
2	Boosting	✓	✓			0.76	88.8%	20.6%
3	OLS Post-LASSO	✓	✓			0.77	89.9%	16.0%
4	Logit Post-LASSO	✓	✓			0.79	89.4%	23.0%
5	Logit Post-LASSO	✓	✓	✓		0.78	89.1%	23.1%
6	Logit Post-LASSO	✓		✓	✓	0.77	87.1%	28.1%

Note: *Individual* indicates that the algorithm has used as inputs all the relevant variables from the student and parent questionnaires. *School* refers to inputs from the teachers, counselor and principal, while *Interaction* indicates that the algorithm includes two-way interaction terms among the top predictors selected by LASSO. *School FE* indicates that school fixed effects were included in the final Logit model.

It is worth mentioning that these algorithms are extremely flexible and can be adapted to different objective function. For instance, if we calibrated the Logit Post-LASSO (Model 4) in order to maximize the area under the ROC curve, we would reach an AUC of 0.81, while maintaining an accuracy of 89.8%, as well as a recall rate of 18.2%. Similarly, If we calibrated the same algorithm to maximize the accuracy rate, we would obtain a similar rate of the one in Table 1 (89.9%), but at the same time the AUC and recall rate would be higher than the ones obtained with the basic model (18.3% and 0.81 compared to 15.2% and 0.80).

[9] We did not actually expected ML to provide perfect predictions. Indeed, as already mentioned, in order to allow schools enough time to identify weak students and target them with appropriate interventions, all predictors were collected in 9th grade. The implicit cost is that the ML algorithms do not take into account all the possible negative shocks affecting educational decisions which may occur between 9th and 12th grade, e.g. unexpected teen pregnancy, health problems, unemployment, and divorce.

[10] It is possible that more sophisticated algorithms may provide even higher performances. However, this would only support the main message of the paper, i.e. that there are big advantages for schools in implementing ML techniques. As discussed in Section 3.3, we have decided to only report results for these three algorithms since there are among the most popular ones. Moreover, their calibration is not extremely time-consuming, thus we avoid the risk that such techniques may be computationally infeasible for schools given their limited technological equipment.

These variations also demonstrate how these high-dimensional techniques can dominate basic models under any performance criterion. Moreover, they show that changing the criterion used to measure performances actually matters and lead to different results, even when there is also one parameter which needs to be selected (the penalization term in LASSO), thus further motivating the need of a theoretically justifies goodness-of-fit measure as discussed in Section 4.

3.5 Pivotal Variables

One way to unpack the black box and understand how boosting obtains the final predictions is to compute the role that each variable has played in the algorithm. As discussed in Friedman (2001) and Schonlau (2005), it is possible to compute the influence of a variable in the boosted regression model estimated in Table 2 (Model 2). This depends on the number of times a variable is chosen across all iterations (trees) and its overall contribution to the log-likelihood function. Such values are then standardized to sum up to 100.

We have therefore looked at the variables which have been selected at least once in the 5-fold estimations. Among the over 1,700 predictors considered, around 140 have been picked by the algorithm to construct a tree. However, around 100 of them have been selected only once, while only 13 of them have been selected 3 or more times. Table 3 lists these 13 predictors along with the number of boosted regression they have been used in, and the 5-fold average influence. [11] Table A1 in the Online Appendix lists the 33 predictors which have been selected at least 2 times.

First of all, it is reassuring to note that there are considerable overlaps between the variables selected by Boosting and the ones used in the heuristic models. Indeed, as highlighted in the previous literature, past academic performances, attendance and school behavior are important predictors. In particular, GPA in 9th grade is always selected and its average influence is rather high. Quite interestingly, gender or ethnicities are not included in this list.

Nevertheless, the list also includes some additional variables which may be useful to improve predictions. ML has indeed been able to detect some indicators which have high predictive power but are often overlooked by practitioners. For instance, not taking any math or science courses in 9th grade plays an important role in the algorithms. This is consistent with the finding in higher education that GPA in math courses is a strong predictor of student retention (Aulck et al., 2016). In line with the previous literature (Bedard and Do, 2005; Schwerdt and West, 2013), transferring school is also related to dropout. Contrary to the wide-spread belief that the ABC (Attendance, Behavior, Course grades) system is able to capture also the impact of family characteristics (Rumberger et al., 2017), number of household members is also often selected, thus highlighting the additional predictive power of household background information. Finally, subjective expectations also matter: the list includes how much the 9th grade is sure of graduating from high school. To summarize, schools correctly use few academic indicators as early warning indicators, but this section has stressed the importance of combining such variables with

[11] The ranking is similar if we sort variables based on the average influence.

additional - carefully selected - predictors and to use advanced techniques to optimally combine them.[12]

Table 3: Variables selected by Boosting

Predictors	Count	Influence
GPA in 9th grade	5	39.7
Born in 1993 (most students were born in 1994-1995)	5	11.2
HSLS:09 Math test score	4	5.9
Whether 9th grader has ever been suspended or expelled	4	5.2
GPA for all academic 9th grade courses	4	2.5
Parent contacted by school about poor attendance more than 4 times	4	2.4
Born in 1992	4	1.9
No science courses taken in 9th grade	3	10.8
No math courses taken in 9th grade	3	4.1
9th grader very sure that he/she will graduate from high school	3	1.5
Credits earned in 9th grade	3	1.3
Number of household members	3	1.1
9th graders has changed schools 7 times since kindergarten	3	0.4

A similar exercise can be conducted with LASSO. In particular, we have looked at the top predictors (around 20-26 in each fold) selected by LASSO to generate the two-way interaction terms in Table 1 Model 6. Among these selected inputs, Table 4 reports the list of variables picked in at least 3 of the 5 folds.

Several variables appears in both Tables 3 and 4: GPA, year of birth, math test score, no math or science course taken in 9th grade, school transfers, attendance, behavior, and expectations about school attainments. It is remarkable that both algorithms select these variables, thus supporting the conclusion regarding their high predictive power. LASSO also frequently selected a few school characteristics, as well as some indicators for parental involvement and parental expectations for student future educational achievements, thus providing policy-makers with additional early-warning indicators with high predictive power.

It may be worth stressing again that these variables are identified by the ML algorithm as important predictors. This does not imply that changing these variables would lead to a reduction of school dropout rates. As also discussed in the introduction, the aim of this paper is to provide precise predictions, not causal inference. This does not reduce the contribution of the paper: both

[12] As also stressed in (Mullainathan and Spiess, 2017), different algorithms and different samples may lead to different variable selections. Indeed, if some variables are highly correlated, then they can substitute each other in predicting school dropout. The actual variables selected depend on the specific finite sample used to train the algorithm. Nevertheless, the aim of this section is to identify top predictors: as long as we obtain accurate predictions, which variables are chosen is irrelevant in this context since we are not assigning a causal interpretation to them. In other words, it is possible that the variables listed in Table 3 may be substituted with other variables, but this would not affect the predictions of the algorithms since – by construction – such variables are highly correlated among themselves.

causality and prediction are relevant in this context since policy-makers are interested in identifying weak students, as well as understanding which variables can be affected to reduce their risk of dropping out.

On a different note, it is necessary to remember that there are several factors which can lead a student to drop out. Therefore, as proved by the results in Table 2, using few indicators cannot match the performances obtained with a larger set of variables.

Table 4: Variables selected by LASSO

Predictors	Count
Born in 1993 (most students were born in 1994-1995)	5
Born in 1995	5
HSLS:09 Math test score	5
No math courses taken in 9^{th} grade	5
No science courses taken in 9^{th} grade	5
GPA in 9^{th} grade	5
9^{th} grader very sure that he/she will graduate from high school	5
Public School	5
Private School	5
Whether 9^{th} grader has ever been suspended or expelled	5
Does not plan to enroll in college after HS	4
Principal reporting student drop out not a problem	4
9^{th} graders has never changed schools since kindergarten	4
Parent reporting no difficulty by 9^{th} grader with behavior problems	4
Parent never contacted by school about poor attendance	4
Parent contacted by school about poor attendance more than 4 times	4
Parent participated in school fundraiser	4
Parent thinks 9^{th} grader will at most attain HS	4
GPA for all academic 9^{th} grade courses	3
9^{th} grader thinks he/she will at most attain HS	3
9^{th} grader did not repeat 2^{nd} grade	3
9^{th} grader spend less than 1h/day on extracurricular activities	3
9^{th} grader was in 9^{th} grade in the previous academic year	3

It is also important to note that most of the predictors in Tables 3-4 are available in administrative data. Therefore, even without collecting additional variables, predictions could be improved by fully leveraging the information contained in the academic transcripts. This may be useful in particular when schools cannot connect their data sets due to privacy issues or prohibitive costs. The latter constraint may be binding especially if these algorithms were applied in developing countries. Even in absence of rich data and with limited resources to expand them,

this section also demonstrates how ML algorithms can be used to identify the key variables from a pilot survey which can then be collected at a larger scale.

An additional advantage of using only administrative data is that they are less manipulable. Indeed, if parents or students were aware that their answers could determine whether or not they are included in a dropout prevention program, they may change the information provided. For instance, they may not truthfully report their expected educational attainments or how many hours they spend playing video games or with friends.[13]

At this point, it is worth noting that the above lists include the math test score administered within the HSLSP:09 survey to all students in 9th grade. However, if we run the same Logit Post-LASSO Model as in Table 2 Model 4 by excluding such variable from the list of potential predictors, we still reach very similar performances (AUC 0.78, accuracy 89.3%, recall 23.2%). Therefore, even if such variable has – as expected – high predictive power, it can be substituted with other predictors in the dataset. Schools are increasingly using entry tests to identify weak students at all educational levels (Shields et al., 2016). Even if this math test score was not designed primarily to detect students at risk of dropping out, the above results suggests that schools can efficiently predict which students are going to drop out without having to rely on expensive tests, but rather by analyzing available individual, family and school characteristics.

3.6 Characteristics Dropout

Students who are predicted to drop out by the ML algorithms but who actually ended up graduating from HS represent a type of error. Indeed, focusing on the predictions obtained from the Logit Post-LASSO algorithm (Table 2 Model 4), around 2% of the students who do not have any reported dropout episode are predicted to dropout. Even if such error rate is relatively small, it still implies that a large number of students targeted by a dropout intervention program does not need such treatment.

However, labelling these students as weak may still be beneficial. Indeed, using information collected in the second follow-up interviews immediately after high school, when individuals were supposed to start their freshman year in college, Table 5 compares average post-secondary outcomes for the 9th graders who are predicted to graduate with those who are predicted to drop out, but only among the students who did not actually drop out from high school in the subsequent years. In other words, following the notation introduced in Section 3.1, Table 5 tests whether students in c_0 are different from those in wr_1.

It is clear from this table that students predicted to drop out who actually graduated from high school are weak students who would gain from some interventions. Indeed, compared to those predicted to graduate, these students "at the margin" of dropping out are 43 percentage points less likely to attend some upper-secondary institution, 18 percentage points more likely to be

[13] However, these data would be manipulable only if individuals were aware of how the prediction of the algorithm would change given the different values of the predictors.

already engaged in starting a family or taking care of their children, and 35 percentage points less likely to have completed the Free Application for Federal Student Aid (FAFSA), a required form to access financial aid. Finally, predicted graduates are also less likely to be working, which is probably due to the higher college attendance.

Therefore, even if these students managed to finish HS against the odds, it does not mean that offering them additional support would not have been beneficial. For instance, additional tutoring or extra classes would have helped them to perform better in higher education. Similarly, counseling would have helped them to choose the appropriate career path. Put differently, targeting students who would otherwise drop out from HS is the main goal of these prevention programs, but including additional weak students could generate further positive effects.

Table 5: Predicted graduates vs predicted dropouts among actual graduates (means)

Additional Outcome	Predicted Graduates	Predicted Dropouts	Difference
Attend post-secondary	0.79	0.36	0.43***
Work	0.52	0.70	-0.18***
Child-care	0.03	0.13	-0.10***
Apply FAFSA	0.70	0.35	0.35***
Observations (Actual Graduates)	14,770	290	

Note: average post-secondary outcomes are compared - among students who actually graduated from high school - for students predicted to graduate or to dropout from high school by the Logit Post-LASSO algorithm (Table 2 Model 4). These variables were collected in 2013, when students were supposed to start their first year of college. FAFSA is the Free Application for Federal Student Aid. All differences are statistically significant at the 1% level, as indicated by the three asterisks next to each figure in the last column.

4. Microeconomic Foundation

So far, we have evaluated model performances by focusing on the recall rate. In this section, we build an economic model to introduce budget considerations and to justify such criterion used in the above analysis. We can start by specifying the optimization problem of the school (or school district officials) in this context: schools want to minimize the expected dropout rate[14] subject to a budget constraint (*BC*).The *BC* takes into account the fact that the individual cost of the dropout prevention program (τ) times the number of students enrolled in the program has to be less or equal to total resources allocated to the program (*B*).

We can now define for each student the probability of dropping out $p(s_i, t_i)$ as a function of her type (s_i) and the treatment (t_i), where the treatment is the dropout prevention program. For simplicity, we can assume that $s_i \in \{0,1\}$. In other words, there are two types of students: students at risk of dropping out (s_i=1) and students not at risk (s_i=0). We would like $p(s_i, t_i)$ to satisfy certain properties:

[14] This objective function is consistent with goals set by federal and state legislation such as Every Student Succeed Act, Race to the Top (U.S. Department of Education, 2009) and the School Progress Report in Philadelphia (District Performance Office, 2017).

$$p(0,t) = 0 \tag{3.1}$$

$$\frac{\partial p(0,t)}{\partial t} = 0 \tag{3.2}$$

$$p(1,t) \geq 0 \tag{3.3}$$

$$\frac{\partial p(1,t)}{\partial t} < 0 \tag{3.4}$$

$$\frac{\partial^2 p(1,t)}{\partial^2 t} > 0 \tag{3.5}$$

Condition (3.1) simply states that students who are not at risk of dropping out have, by definition, a zero probability of dropping out given any treatment. Similarly, condition (3.2) ensures that the probability of dropping out for students not at risk is not affected by the level of treatment. Condition (3.3) means that the probability of dropping out for students at risk is non-negative. Condition (3.4) makes clear that treatment is effective: more intense treatment decreases the probability of dropping out for weak students. Finally, condition (3.5) implies decreasing returns to scale, thus it is optimal to allocate resources equally among weak students.

However, schools do not directly observe students at risk, but rather only a signal, i.e. a predicted probability of dropping out provided by an algorithm. Therefore, using the confusion matrix introduced in Section 3.1, the school optimization problem becomes:

$$\min n_1[(1-\varphi)p(1,0) + \varphi p(1,t)]$$
$$s.t.\ \tau t[wr_1 + c_1] \leq B$$

Where the objective function is the weighted sum of the number of students at risk of dropping out which are not treated plus those who are treated, each multiplied by the probability of dropping out given the treatment. On the other hand, the cost of the program in the budget constraint depends on the students which have been - both correctly and incorrectly - assigned to the treatment.

In order to obtain a closed-form expression, we can add two assumptions. First, let $t_i \in \{0,1\}$, so that students can only be included or excluded from the dropout prevention program. This is realistic in a setting in which a program has already been designed and schools are only required to identify the weakest students who need to be included in such program. In other words, individual, family and school characteristics are used to identify s_i, i.e. to find out who are the weak students, thus providing a signal to schools. Condition (3.5) is no longer required. Given this additional assumption, we can impose the following functional form:

$$p(s_i, t_i) = (1 - t_i)s_i$$

Note that this functional form satisfies conditions (3.1)-(3.4). From this, it follows that the objective function becomes (excluding the constant n_1):

$$\min(1-\varphi) * 1 + \varphi * 0$$
$$s.t.\ \tau[wr_1 + c_1] \leq B$$

Which is equivalent to maximize the Recall rate subject to a BC. We have therefore derived an economic justification to use such criterion when tuning the ML algorithms using CV and when comparing performances among them.

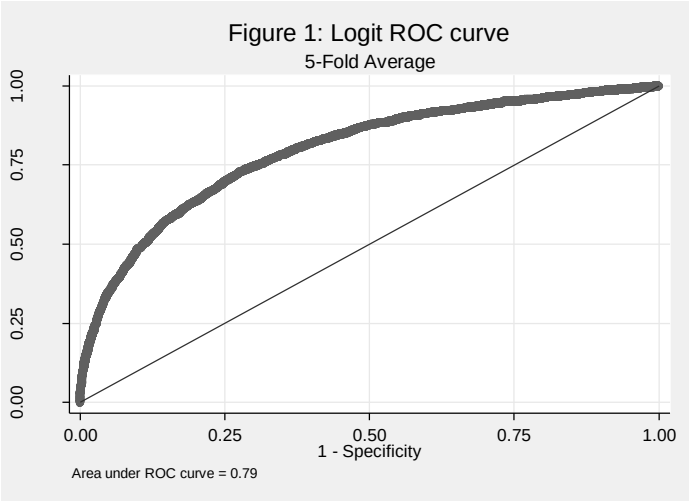
We should stress that using the Recall rate as criterion can be justified also by imposing different functional forms on $p(s_i, t_i)$. Indeed, if we take into account the curvature imposed by the positive second derivative of $p(s_i, t_i)$ (condition 3.5), we could assume the following functional form:

$$p(s,t) = \frac{s}{1+t}$$

Which would lead to an equivalent optimization problem:

$$\min 1 - \frac{\varphi}{2}$$
$$s.t.\ \tau[wr_1 + c_1] \leq B$$

A straightforward implementation of the above procedure can be applied to the Logit model discussed in Section 3.2 (Table 1 Model 1). Previously, we used 0.5 as threshold. However, we can change this parameter to maximize the recall rate in the CV sample while respecting the budget constraint. This can be interpreted as choosing a point in the ROC curve depicted in Figure 1[15]: we want to be as high as possible on the y-axis, but the selected point cannot be too much on the right of the x-axis otherwise the program exceeds the resources available. Indeed, after estimating the individual probability of drop out for each student, the ROC curve is obtained by letting the probability threshold used to divide students between predicted graduates and dropouts to vary between zero and one, and by then computing the resulting sensitivity and specificity for each cutoff. In the bottom-left corner, specificity is one, that is the algorithm perfectly predicts those who are going to graduate, but sensitivity is zero, thus the algorithm does not identify any of the students who end up dropping out. On the other hand, in the top-right corner, sensitivity is one, thus the algorithm perfectly predicts those who are going to drop out, but specificity is zero, meaning that none of the graduating students are identified as HS graduates. Therefore, instead of using the area under the ROC curve as main criterion to compare algorithms as in Bowers et al. (2013), we propose a theoretical model to justify the selection of the optimal point on the ROC curve.



[15] The Online Appendix provides a detailed explanation of how Figure 1 and Table 6 in this section have been computed.

Table 6 shows how the optimal accuracy and recall rates change as we vary the cost per student and the overall budget of the program.[16] As a result, following this procedure policy-makers can choose the most efficient algorithm and tune its parameters in order to treat as many weak students as possible subject to their budget constraints. It is also worth noting that, thanks the low variability of the Logit estimates between in-sample and out-of-sample (because of the small number of predictors compared to the sample size), the actual costs incurred by the school - that is the overall expenditure obtained using the test sample - is similar to the planned cost. In other words, the advantage of using an algorithm with low variance is that there is a lower risk that the cost of a dropout prevention program does eventually exceed the resources initially allocated to it.

Table 6: Optimal Threshold – 5-Fold Average

		Overall Budget		
Cost per student		1,000	10,000	100,000
10	Actual Cost	970	9,714	42,244
	Threshold	0.53	0.14	0.01
	Accuracy	89.9%	79.8%	12.2%
	Recall	13.9%	61.6%	99.7%
100	Actual Cost	1,020	9,700	97,140
	Threshold	0.84	0.53	0.14
	Accuracy	89.4%	89.9%	79.8%
	Recall	2%	13.9%	61.6%
500	Actual Cost	1,000	9,700	98,400
	Threshold	0.93	0.77	0.40
	Accuracy	89.3%	89.6%	89.6%
	Recall	0.4%	3.6%	22.9%

5. Clustering Predicted Dropouts

Identifying weak students is only the first step. Next, schools have to design the appropriate programs for them. However, as also stressed in Bowers and Sprott (2012), these students do not represent a homogeneous groups and they may need different treatments. For instance, students who are struggling academically may benefit from tutoring or summer classes, while counseling may be more effective for students with discipline issues or problems at home. Income inequality may also play a role for individuals from low socio-economic background (Kearney and Levine, 2016). In other words, this section acknowledges that HS dropout is a multidimensional issue: different factors may lead students to halt their education. This is similar in spirit to the multidimensional approach advocated in poverty studies (Alkire and Foster, 2011). Therefore, it shows how to divide the students predicted to dropout into different subgroups using unsupervised machine learning.

The starting point is the predictions obtained using the Logit Post-LASSO algorithm in Table 2 (Model 4). In line with the results in Section 3.5, we have then used the same predictors selected

[16] It is worth remembering that in the CV (as well as Test) sample there are around 4,290 students and 460 dropouts.

by this algorithm at least in 3 of the 5 folds (Table 4) and we have divided the students predicted to dropout into different groups by means of a hierarchical clustering algorithm. As explained in the Online Appendix, the Caliński and Harabasz pseudo-F index and the Duda-Hart Je(2)/Je(1) index with associated pseudo-T^2 can help analysts to select the best number of groups, four in this case. Table 7 shows the summary statistics for these predicted dropouts. For comparison, the second column also includes the summary statistics for the students who are predicted to graduate.

There are some similarities between these four groups. All these students had very low academic performances in terms of GPA and math test scores. Moreover, almost all of them were attending public schools, and their principals were more likely than others to report that student dropout was an issue in their school. Despite these similarities, there are several striking differences among these clusters, which thus suggest that they indeed require different kinds of support.[17]

Group 1 is mainly composed by individuals with low attendance, behavioral issues, and lack of parental involvement. On the other hand, students in Group 2 were older than the usual 9th grader, thus indicating that they had already repeated a grade. They were also characterized by very low expectations: both the students and their parents were more likely to believe that they would at most graduate from high school. Group 3 includes mainly students who had been suspended or expelled, with frequent attendance issues, who were already repeating 9th grade, and who were not taking any math or science course.

Finally, students in Group 4 are rather peculiar: they were quite sure that they would have graduated from high school, and this belief was shared by their parent. They were also planning to enroll in college, they had good attendance records, and their parents were involved in their education. Nevertheless, they had low academic performances, and many of them were already in 9th grade in the previous academic year. Therefore, this result stresses again the importance of not pooling together all students at risk of dropping out: place these students in a classroom side by side with students from the other groups may actually results in negative externalities.

[17] It is also worth mentioning that, since the recall rate is not 100%, all these groups contains students who actually graduated from high school even if they were predicted not to. Nevertheless, these misclassified students are not concentrated in one cluster only: each group contains a similar percentage of correctly and incorrectly predicted dropouts.

Table 7: Clustering

Predictors	No Dropout	Group 1	Group 2	Group 3	Group 4
Born in 1993 (most students were born in 1994-1995)	0.03	0.22	0.63	0.13	0.65
Born in 1995	0.58	0.30	0.01	0.07	0.02
HSLS:09 Math test score	0.47	0.31	0.31	0.32	0.33
No math courses taken in 9^{th} grade	0.04	0.26	0.19	0.40	0.10
No science courses taken in 9^{th} grade	0.06	0.31	0.25	0.59	0.14
GPA in 9^{th} grade	0.63	0.25	0.27	0.21	0.31
9^{th} grader very sure that he/she will graduate from high school	0.84	0.45	0.33	0.64	0.77
Public School	0.81	0.99	0.99	1.00	0.98
Private School	0.12	0.00	0.00	0.00	0.00
Whether 9^{th} grader has ever been suspended or expelled	0.07	0.20	0.49	0.92	0.49
Does not plan to enroll in college after HS	0.44	0.79	0.96	0.87	0.59
Principal reporting student drop out not a problem	0.28	0.04	0.04	0.05	0.08
9^{th} graders has never changed schools since kindergarten	0.34	0.08	0.37	0.17	0.13
Parent reporting no difficulty by 9^{th} grader with behavior problems	0.64	0.20	0.49	0.29	0.59
Parent never contacted by school about poor attendance	0.62	0.20	0.30	0.17	0.74
Parent contacted by school about poor attendance more than 4 times	0.01	0.09	0.18	0.39	0.25
Parent participated in school fundraiser	0.39	0.11	0.10	0.14	0.39
Parent thinks 9^{th} grader will at most attain HS	0.05	0.15	0.59	0.34	0.04
GPA for all academic 9^{th} grade courses	0.60	0.22	0.25	0.19	0.28
9^{th} grader thinks he/she will at most attain HS	0.11	0.35	0.73	0.32	0.12
9^{th} grader did not repeat 2^{nd} grade	0.05	0.04	0.81	0.79	0.77
9^{th} grader spend less than 1h/day on extracurricular activities	0.30	0.60	0.77	0.46	0.31
9^{th} grader was in 9^{th} grade in the previous academic year	0.04	0.23	0.33	0.43	0.39
Observations	20,340	630	110	120	100

Note: All variables has been rescales between 0 and 1

6. Conclusions

This paper shows how schools can promptly identify students at risk of dropping out by using available high dimensional data jointly with ML techniques. One of the advantage of these early predictions is that in this way counselors and teachers may also suggest to weak students to consider vocational careers (Goux et al., 2016). More generally, this paper proves that Big Data and ML can be fruitfully applied in education and lead to improved school performances thanks to an efficient use of the available information.

From a policy perspective, this contribution could lead to a substantial reduction in dropout rates if schools used the proposed algorithm to target weak students and draw from the existing literature to identify effective programs to help them. Although using few indicators may be attractive, this paper stresses that this approach leads to extremely unreliable predictions. Schools have several more data available to them, and their dimension is increasing exponentially over time thanks to new technologies: ML can help practitioners to efficiently use this new information. Data analysts can easily develop a user-interface to automatically implement ML algorithms (Aguiar et al., 2015; Knowles, 2015), thus allowing teachers and administrators to readily identify students at-risk without having to rely on few early warning indicators for the sake of simplicity.

Even when schools have limited records - which is often the case in developing countries - ML extract all the prediction power of the available data. Moreover, schools in these countries could use the results from the U.S. or from pilot studies to understand which variables have a bigger role and thus are worth collecting at a national level.

Furthermore, we have showed not only that supervised ML can improve school predictions, but also that unsupervised ML can identify sub-populations among the weak students. Therefore, schools would be able to design the appropriate program for each group by understanding their peculiarity and the key factors which are associated with their low performances. In other words, rather than offering the same intervention to all students in all schools, policymakers can exploit these algorithms to personalize the treatment which each cluster of students in the school requires in order to improve their academic performances.

From an economic point of view, this paper contributes to the ML literature by constructing a microeconomic model to justify the criterion used in evaluating the performances of the algorithms. This is rather important in a context in which there is no clear measurement and practitioners tend to (quite arbitrarily) choose among a large set of possible performance evaluations.

Another way to justify the focus of this paper on prediction is to view it as a targeting application. Let's assume that there are two types of students, those at risk and those not at risk of dropping out, and that there is an effective treatment which can be provided by schools and which has a homogeneous impact on weak students. In other words, we assume that there is a dropout prevention program which is able to equally reduce the probability of dropping out for

all treated struggling students.[18] High-dosage tutoring is an example of such a policy able to help these students (Fryer, 2017). The necessary pre-condition to implement this program is to identify the students who need the treatment, i.e. those at risk of not graduating from HS. This is the context in which the algorithms presented in this paper can be successfully applied: ML can efficiently use the information available to schools in order to identify students which can be included in the program. Schools need to know if their students belong to the "not at risk" or "at risk category", and ML can provide them an accurate signal of student type for each individual.

More generally, supervised ML can be used in the first stage to identify students who are more at risk of dropping out among the whole student population, while unsupervised ML can divide these students into subgroups, and then scarce and expensive human resources can be invested to decide the best intervention for these restricted set of students. Therefore, even if current ML techniques are designed to provide accurate predictions, but they are often inappropriate to optimally allocate resources (Athey, 2017), they can provide complementary tools for causal inference.

[18] Note that we do not require homogeneous treatment for the whole population, but only for the weak students. In fact, the treatment may be completely ineffective for students who have high probability of graduating from HS.

References

Adelman, M., Haimovich, F., Ham, A., Vazquez, E., 2017. Predicting School Dropout with Administrative Data: New Evidence from Guatemala and Honduras. World Bank Policy Res. Work. Pap. 8142.

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B., Addison, K.L., 2015. Who, when, and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time. Proc. Fifth Int. Conf. Learn. Anal. Knowl. - LAK '15 93–102.

Alkire, S., Foster, J., 2011. Counting and multidimensional poverty measurement. J. Public Econ. 95, 476–487.

Athey, S., 2017. Beyond prediction: Using big data for policy problems. Sci. Mag. 355, 483–485.

Aulck, L., Velagapudi, N., Blumenstock, J., West, J., 2016. Predicting Student Dropout in Higher Education (No. arXiv:1606.06364), arXiv.

Bauer, E., Kohavi, R., Chan, P., Stolfo, S., Wolpert, D., 1999. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Mach. Learn. 36, 105–139.

Bedard, K., Do, C., 2005. Are Middle Schools More Effective?: The Impact of School Structure on Student Outcomes. J. Hum. Resour. 40, 660–682.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. J. Econ. Perspect. 28, 29–50.

Bowers, A.J., Sprott, R., 2012. Why Tenth Graders Fail to Finish High School: A Dropout Typology Latent Class Analysis. J. Educ. Students Placed Risk 17, 129–148.

Bowers, A.J., Sprott, R., Taff, S.A., 2013. Do We Know Who Will Drop Out?: A Review of the Predictors of Dropping out of High School: Precision, Sensitivity, and Specificity. High Sch. J. 96, 77–100.

Carnevale, A.P., Smith, N., Strohl, J., 2013. Recovery - Job Growth and Education Requirements through 2020. Washington, D.C.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., Mullainathan, S., 2016. Productivity and Selection of Human Capital with Machine Learning. Am. Econ. Rev. Pap. Proc. 106, 124–127.

De Witte, K., Cabus, S., Thyssen, G., Groot, W., Van Den Brink, H.M., 2013. A critical review of the literature on school dropout. Educ. Res. Rev. 10, 13–28.

Deussen, T., Hanson, H., Bisht, B., 2017. Are Two Commonly Used Early Warning Indicators Accurate Predictors of Dropout for English Learner Students? Evidence from Six Districts in Washington State, Regional Educational Laboratory Northwest. Washington, D.C.

District Performance Office, 2017. 2015-2016 School Progress Report. Philadelphia.

Ekowo, M., Palmer, I., 2016. The Promise and Peril of Predictive Analytics in Higher Education. Washington, D.C.

Friedman, J., 2001. Greedy Function Approximation : A Gradient Boosting Machine. Ann. Stat. 29, 1189–1232.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive Logistic Regression: A Statistical View of Boosting. Ann. Stat. 28, 337–407.

Fryer, R.G., 2017. The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments, in: Duflo, E., Banerjee, A. (Eds.), Handbook of Field Experiments. North Holland, Amsterdam, pp. 95–322.

Giannone, D., Lenza, M., Primiceri, G.E., 2017. Economic Predictions with Big Data: The

Illusion of Sparsity.
Goux, D., Gurgand, M., Maurin, E., 2016. Adjusting Your Dreams? High School Plans and Dropout Behaviour. Econ. J. 1–22.
Guenther, N., Schonlau, M., 2016. Support Vector Machines. Stata J. 16, 917–937.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edi. ed, Springer Series in Statistics. Springer.
Heckman, J.J., Humphries, J.E., Mader, N.S., 2011. The GED, in: Hanushek, E.A., Machin, S., Woessmann, L. (Eds.), Handbooks in Economics: Economics of Education. North-Holland, pp. 423–484.
Heckman, J.J., LaFontaine, P.A., 2010. The American High School Graduation Rate: Trends and Levels. Rev. Econ. Stat. 92, 244–262.
Heckman, J.J., Rubinstein, Y., 2001. The Importance of Noncognitive Skills: Lessons from the GED Testing Program. Am. Econ. Rev. Pap. Proc. 91, 145–149.
IES, 2016. Public High School Graduation Rate Reaches New High, but Gaps Persist. Washington, D.C.
Ingels, S.J., Pratt, D.J., Herget, D., Bryan, M., Fritch, L.B., Ottem, R., Rogers, J.E., Wilson, D., 2015. High School Longitudinal Study of 2009 (HSLS : 09) 2013 Update and High School Transcript Data File Documentation. Washington, DC.
Ingels, S.J., Pratt, D.J., Herget, D.R., Burns, L.J., Dever, J.A., Ottem, R., Rogers, J.E., Jin, Y., Leinwand, S., 2011. High School Longitudinal Study of 2009 (HSLS:09). Base-Year Data File Documentation. Washington, DC.
Ingels, S.J., Pratt, D.J., Herget, D.R., Dever, J.A., Fritch, L.B., Ottem, R., Rogers, J.E., Kitmitto, S., Leinwand, S., 2014. High School Longitudinal Study of 2009 (HSLS:09) Base Year to First Follow-Up Data File Documentation. Washington, DC.
Kearney, M.S., Levine, P.B., 2016. Income Inequality, Social Mobility, and the Decision to Drop Out of High School. Brookings Pap. Econ. Act. Spring, 333–380.
Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction Policy Problems. Am. Econ. Rev. Pap. Proc. 105, 491–495.
Knowles, J., 2015. Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. JEDM - J. Educ. Data Min. 7, 1–52.
Luca, M., Kleinberg, J., Mullainathan, S., 2016. Algorithms Need Managers, Too. Harv. Bus. Rev. 104, 96–101.
Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., de Mendonça, A., 2011. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res. Notes 4, 299.
Mullainathan, S., Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. J. Econ. Perspect. 31, 87–106.
Murnane, R.J., 2013. U.S. High School Graduation Rates: Patterns and Explanations. J. Econ. Lit. 51, 370–422.
Ng, A., 2016. Machine Learning [WWW Document]. Coursera. URL https://www.coursera.org/learn/machine-learning (accessed 1.1.16).
O'Cummings, M., Therriault, S.B., 2015. From Accountability to Prevention : Early Warning Systems Put Data to Work for Struggling Students. Washington, D.C.
OECD, 2016. Education at a Glance 2016: OECD Indicators. OECD Publishing, Paris.

Oreopoulos, P., 2007. Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling. J. Public Econ. 91, 2213–2229.
Oreopoulos, P., Salvanes, K.G., 2011. Priceless: The Nonpecuniary Benefits of Schooling. J. Econ. Perspect. 25, 159–184.
Rumberger, R., Addis, H., Allensworth, E., Balfanz, R., Bruch, J., Dillon, E., Duardo, D., Dynarski, M., Furgeson, J., Jayanthi, M., Newman-Gonchar, R., Place, K., Tuttle, C., 2017. Preventing Dropout in Secondary Schools. Washington, D.C.
Rumberger, R.W., Lim, S.A., 2008. Why students drop out of school: A review of 25 years of research. UC Santa Barbara.
Sara, N.-B., Halland, R., Igel, C., Alstrup, S., 2015. High-School Dropout Prediction Using Machine Learning : A Danish Large-scale Study. ESANN 2015 proceedings, Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn. 22–24.
Schonlau, M., 2005. Boosted regression (boosting): An introductory tutorial and a Stata plugin. Stata J. 5, 330–354.
Schwerdt, G., West, M.R., 2013. The impact of alternative grade configurations on student outcomes through middle and high school. J. Public Econ. 97, 308–326.
Shields, K.A., Cook, K.D., Greller, S., 2016. How Kindergarten Entry Assessments Are Used in Public Schools and How They Correlate with Spring Assessments. Washington, D.C.
Subrahmanian, V.S., Kumar, S., 2017. Predicting human behavior: The next frontiers. Sci. Mag. 355, 489.
U.S. Department of Education, 2009. Race to the Top Program: Executive Summary.
Varian, H.R., 2014. Big Data : New Tricks for Econometrics. J. Econ. Perspect. 28, 3–28.
Witten, D.M., Tibshirani, R., 2010. A Framework for Feature Selection in Clustering. J. Am. Stat. Assoc. 105, 713–726.
Zajacova, A., 2012. Health in working-aged Americans: Adults with high school equivalency diploma are similar to dropouts, not high school graduates. Am. J. Public Health 102, 284–290.

Online Appendix

A.1 Variable descriptions

This appendix contains the detailed description of all the variables used in the empirical sections.

A.1.1 Basic Model Predictors

Female is an indicator variable equal to one if the student's sex is female, zero if the student's sex is male.

Ethnicity is a series of indicator variables to specify the student's ethnicity. White has been used as comparison group. Native Hawaiian, Pacific Islander, American Indian, and Alaska Native (non-Hispanic) have been grouped together as "Other Ethnicity".

Language indicates that the language the student first learned to speak was a non-English language only. The comparison group comprises students whose first language is English only, or English and a non-English language equally. This information has been obtained from the second follow-up interviews because of the lower number of missing values than the baseline survey.

Region is a series of indicator variables recording the school geographic region (Midwest, South, West). Northeast has been used as comparison group.

Urbanicity is a series of indicator variables recording the school locale (urbanicity). The categories are suburb, town, rural. City has been used as comparison group.

School climate is a scale of the administrator's assessment of the school climate. It has been created through weighted principal component factor analysis and standardized with a mean of zero and standard deviation of one. The inputs of this scale were the answer to the following questions:

> To the best of your knowledge how often do the following types of problems occur at your high school?

- Physical conflicts among students
- Robbery or theft
- Vandalism
- Student use of illegal drugs while at school
- Student use of alcohol while at school
- The sale of drugs on the way to or from school or on school grounds
- Student possession of weapons
- Physical abuse of teachers
- Student racial tensions
- Student bullying
- Student verbal abuse of teachers
- Student in-class misbehavior
- Student acts of disrespect for teachers
- Student gang activities

Only respondents who provided a full set of responses were assigned a scale value. Additional information on this scale is available in Chapter 5 of Ingels et al. (2011).

Poverty is an indicator variable reporting whether the student lived in a household whose income is below the 2011 poverty threshold set by the U.S. Census Bureau. This information has been obtained from the second follow-up interviews because of the lower number of missing values than the baseline survey.

Number of HH Members records the number of household member. This variable has been top-coded as 19 for households with more than 18 members. This information has been obtained from the second follow-up because of the lower number of missing values than the baseline survey.

No mother/father are two indicator variables reporting whether there was no biological/adoptive/step-mother/father in the household. This information has been obtained from the second follow-up interviews because of the lower number of missing values than the baseline survey.

Mother/father education are two indicator variables recording whether the educational level of the biological/adoptive/step-mother or father is lower than high school (diploma, GED or alternative high school credential.). Both these variables were imputed to zero if there was no biological/adoptive/step-mother or father in the household. This information has been obtained from the second follow-up interviews because of the lower number of missing values than the baseline survey.

Mother/father employment are two indicator variables recording whether the biological/adoptive/step-mother or father was working part-time (less than 35 hours/week) or full-time at the time of the survey. Both these variables were imputed to zero if there was no biological/adoptive/step-mother or father in the household. This information has been obtained from the second follow-up interviews because of the lower number of missing values than the baseline survey.

Math Score (9th grade) provides a norm-referenced measurement of achievement, that is, an estimate of achievement relative to the population (Fall 2009 9th graders) as a whole. It provides information on status compared to peers. This feature is the main difference from the IRT-estimated percent-correct scores, which represent status with respect to achievement on a particular criterion set of test items. Such mathematical assessment focused on algebra skills, reasoning, and problem solving. It was developed specifically for the HSLS:09 and was administered to students in 9th grade and 11th grade. See Chapter 2 in Ingels et al. (2011) and Ingels et al. (2014) for more information about the test and the derivation of the normalized scores.

GPA (9th grade) is extracted from the student's high school transcript and it contains the student's GPA in 9th grade.

Credits (9th grade) are extracted from the student's high school transcript and it contains the student's total Carnegie credits earned in 9th grade. A Carnegie unit is equivalent to a one-year academic course taken one period a day, five days a week.

Repeated 9th grade is an indicator variable equal to one if the student was in 9th grade in the previous academic year (2008/2009), zero if she was in 7th grade, 8th grade, or an ungraded program.

Attendance and Suspension records the number of times the student reported skipping class or being put on an in-school suspension in the previous 6 months. This information has been obtained from the second follow-up interviews because it is not available in the baseline survey.

A.1.2 ML Predictors

The following list reports the original name of all the variables which has been used as inputs. Variables starting with X refers to imputed variables, S refers to variables obtained from the students, P from the parents, M from the math teachers, N from the science teachers, A from the principals, C from the counselors. Additional information is available in the HSLS:09 Codebook. For a few variables (those whose second letter is not a 1), information has been obtained from the first or second follow-up interviews because of the lower number of missing values than the baseline survey.

For each categorical variable, we have generated a series of indicator variables, one for each possible value of the categorical variable. Missing values and unit-non-response have also been considered as two distinct values. In other words, if a variable Y had three possible values (A, B, C) and some missing values, we have generated one indicator variable equal to one when Y was equal to A, zero when it was equal to B, C or missing. Similar indicators variables have been constructed for the values B and C, as well as for missing values. For continuous variables, we have imputed missing values to zero and generated indicator variables to report for which observations the variables were imputed. All these variables have been used as predictors in the ML algorithms.

1. X1SEX: Student's sex
2. X1RACE: Student's race/ethnicity-composite
3. X2HISPTYPE: Student's Hispanic/Latino/Latina subgroup-composite
4. X2ASIANTYPE: Student's Asian subgroup-composite
5. X1NATIVELANG: Student's native language
6. X2DUALLANG: Student dual-first language indicator
7. X2STDOB: Student's date of birth (YYYYMM)
8. X1TXMTSCOR: Mathematics standardized theta score
9. X2MOMREL: Mother/female guardian's relationship to 9th grader
10. X2MOMEDU: Mother's/female guardian's highest level of education
11. X2MOMEMP: Mother/female guardian's employment status
12. X2MOMOCC2: Mother/female guardian's current/most recent occupation: 2-digit ONET
13. X1MOMRACE: Mother's race/ethnicity
14. X2DADREL: Father/male guardian's relationship to 9th grader
15. X2DADEDU: Father's/male guardian's highest level of education
16. X2DADEMP: Father/male guardian's employment status
17. X2DADOCC2: Father/male guardian's current/most recent occupation: 2-digit ONET
18. X1DADRACE: Father's race/ethnicity
19. X2HHNUMBER: Number of 2012 household members
20. X2FAMINCOME: Total family income from all sources 2011
21. X1STU30OCC2: Student expected occupation at age 30: 2-digit ONET
22. X1STUEDEXPCT: How far in school 9th grader thinks he/she will get
23. X1PAREDEXPCT: How far in school parent thinks 9th grader will go
24. X1PQLANG: Parent questionnaire language (English v. Spanish)

25. X1CONTROL: School control (public, Catholic, other private)
26. X1LOCALE: School locale (urbanicity)
27. X1CENDIV: School census geographic division
28. X1GRADESPAN: Grade span of school-administrator questionnaire
29. X3THIMATH9: Highest level mathematics course taken - ninth grade
30. X3TWHENALG1: When student took algebra I
31. X3THISCI9: Highest level science course taken - ninth grade
32. X3TCRED9TH: Credits earned in: ninth grade
33. X3TAGPA09 : GPA for all academic 9th grade courses
34. X3TGPA9TH: GPA: ninth grade
35. S1GRD0809: Grade 9th grader was in last year (2008-09)
36. S1SCH0809: Whether 9th grader attended a different school last year (2008-09)
37. S1M8GRADE: Final grade in 9th grader's most advanced 8th grade math course
38. S1S8GRADE: Final grade in 9th grader's most advanced 8th grade science course
39. S1MPERSON1 : 9th grader sees himself/herself as a math person
40. S1MPERSON2 : Others see 9th grader as a math person
41. S1MUNDERST : How often 9th grader thinks he/she really understands math assignments
42. S1MTESTS: 9th grader confident can do excellent job on fall 2009 math tests
43. S1MTEXTBOOK : 9th grader certain can understand fall 2009 math textbook
44. S1MSKILLS: 9th grader certain can master skills in fall 2009 math course
45. S1MASSEXCL: 9th grader confident can do excellent job on fall 2009 math assignments
46. S1SPERSON1: 9th grader sees himself/herself as a science person
47. S1SPERSON2: Others see 9th grader as a science person
48. S1SUNDERST: How often 9th grader thinks he/she really understands science assignments
49. S1STESTS: 9th grader confident can do excellent job on fall 2009 science tests
50. S1STEXTBOOK: 9th grader certain can understand fall 2009 science textbook
51. S1SSKILLS: 9th grader certain can master skills in fall 2009 science course
52. S1SASSEXCL: 9th grader confident can do excellent job on fall 09 science assignments
53. S1SAFE: 9th grader feels safe at school
54. S1PROUD: 9th grader is proud to be part of his/her school
55. S1TALKPROB: 9th grader has teacher/adult in school he/she can talk to about problems
56. S1SCHWASTE: 9th grader feels that school is often a waste of time
57. S1GOODGRADES: Getting good grades is important to 9th grader
58. S1NOHWDN: How often 9th grader goes to class without their homework done
59. S1NOPAPER: How often 9th grader goes to class without pencil or paper
60. S1NOBOOKS: How often 9th grader goes to class without books
61. S1LATE: How often 9th grader goes to class late
62. S1FAVSUBJ: 9th grader's favorite school subject
63. S1LEASTSUBJ: 9th grader's least favorite school subject
64. S1PAYOFF: 9th grader thinks studying in school rarely pays off later with good job
65. S1GETINTOCLG: 9th grader thinks even if he/she studies he/she won't get into college
66. S1AFFORD: 9th grader thinks even if he/she studies family can't afford college
67. S1WORKING: 9th grader thinks working is more important for him/her than college
68. S1MOMTALKCLG: 9th grader talked to mother about going to college
69. S1DADTALKCLG: 9th grader talked to father about going to college
70. S1FRNDTLKCLG: 9th grader talked to friends about going to college

71. S1TCHTALKCLG: 9th grader talked to teacher about going to college
72. S1CNSLTLKCLG: 9th grader talked to school counselor about going to college
73. S1MOMTALKJOB: 9th grader talked to mother about adult jobs/careers
74. S1DADTALKJOB: 9th grader talked to father about adult jobs/careers
75. S1FRNDTLKJOB: 9th grader talked to friends about adult jobs/careers
76. S1TCHTALKJOB: 9th grader talked to teacher about adult jobs/careers
77. S1CNSLTLKJOB: 9th grader talked to school counselor about adult jobs/careers
78. S1MOMTALKPRB: 9th grader talked to mother about personal problems
79. S1DADTALKPRB: 9th grader talked to father about personal problems
80. S1FRNDTLKPRB: 9th grader talked to friends about personal problems
81. S1TCHTALKPRB: 9th grader talked to teacher about personal problems
82. S1CNSLTLKPRB: 9th grader talked to school counselor about personal problems
83. S1FRNDGRADES: 9th grader's closest friend gets good grades
84. S1FRNDSCHOOL: 9th grader's closest friend is interested in school
85. S1FRNDCLASS: 9th grader's closest friend attends classes regularly
86. S1FRNDCLG: 9th grader's closest friend plans to go to college
87. S1TEFRNDS: Time/effort in math/science means not enough time with friends
88. S1TEACTIV: Time/effort in math/science means not enough time for extracurricular
89. S1TEPOPULAR: Time/effort in math/science means 9th grader won't be popular
90. S1TEMAKEFUN: Time/effort in math/science means people will make fun of 9th grader
91. S1ENGCOMP: How 9th grader compares males and females in English or language arts
92. S1MTHCOMP: How 9th grader compares males and females in math
93. S1SCICOMP: How 9th grader compares males and females in science
94. S1HRMHOMEWK: Hours spent on math homework/studying on typical school day
95. S1HRSHOMEWK: Hours spent on science homework/studying on typical school day
96. S1HROTHHOMWK: Hours spent on other homework/studying on typical school day
97. S1HRACTIVITY: Hours spent on extracurricular activities on typical school day
98. S1HRWORK: Hours spent working for pay on typical school day
99. S1HRFAMILY: Hours spent with family on typical school day
100. S1HRFRIENDS: Hours spent hanging out with friends on typical school day
101. S1HRTV: Hours spent watching television or movies on typical school day
102. S1HRVIDEO: Hours spent playing video games on typical school day
103. S1HRONLINE: Hours spent chatting or surfing online on typical school day
104. S1MYRS: Number of years of math coursework 9th grader expects to take in HS
105. S1APCALC: 9th grader plans to enroll in an Advanced Placement (AP) calculus course
106. S1IBCALC: 9th grader plans to enroll in International Baccalaureate (IB) calculus
107. S1SYRS: Number of years of science coursework 9th grader expects to take in HS
108. S1APS: 9th grader plans to enroll in an Advanced Placement (AP) science course
109. S1IBSCI: 9th grader plans to enroll in International Baccalaureate (IB) science
110. S1PLAN: 9th grader has put together an education plan and/or career plan
111. S1PSAT: 9th grader has taken or plans to take the PSAT
112. S1SAT: 9th grader has taken or plans to take the SAT
113. S1ACT: 9th grader has taken or plans to take the ACT
114. S1AP: 9th grader has taken/plans to take an Advanced Placement (AP) test
115. S1IBTEST: 9th grader has taken/plans to take International Baccalaureate (IB) test
116. S1SUREHSGRAD: How sure 9th grader is that he/she will graduate from high school

117. S1SURECLG: How sure 9th grader is that he/she will go to college to pursue a BA/BS
118. S1ABILITYBA: 9th grader thinks he/she has the ability to complete a Bachelor's degree
119. S1BAAGE30: 9th grader would be disappointed if he/she didn't have a BA/BS by age 30
120. S1FYAA: 9th grader plans to enroll in Associate's program in 1st year after HS
121. S1FYBA: 9th grader plans to enroll in Bachelor's program in 1st year after HS
122. S1FYLICENSE: 9th grader plans to obtain license or certificate in 1st year after HS
123. S1FYAPPR: 9th grader plans to attend apprenticeship program in 1st year after HS
124. S1FYMILITARY: 9th grader plans to join the armed services in 1st year after HS
125. S1FYJOB: 9th grader plans to get a job in 1st year after HS
126. S1FYFAMILY: 9th grader plans to start a family in 1st year after HS
127. S1FYTRAVEL: 9th grader plans to travel in 1st year after HS
128. S1FYVOLUN: 9th grader plans to volunteer or do missionary work in 1st year after HS
129. S1FYNOTSURE: 9th grader does not know what he/she will do in 1st year after HS
130. S1ESTIN: Estimate of tuition and mandatory fees at public in-state 4-year college
131. P1HHLT18: Number of household residents less than 18 years of age
132. P1HHTIME: How much of the time 9th grader lives with respondent
133. P1HSSIB: 9th grader has sibling who attends/attended his/her HS in last 5 years
134. P1OLDERSIB: Number of older siblings
135. P1USBORN9: Whether student was born in the U.S.
136. P1OWNHOME: Home is owned, rented or other arrangement
137. P1REPEATGRD: Ninth grader has repeated a grade
138. P1REPEATGK: Ninth grader repeated kindergarten
139. P1REPEATG1: Ninth grader repeated 1st grade
140. P1REPEATG2: Ninth grader repeated 2nd grade
141. P1REPEATG3: Ninth grader repeated 3rd grade
142. P1REPEATG4: Ninth grader repeated 4th grade
143. P1REPEATG5: Ninth grader repeated 5th grade
144. P1REPEATG6: Ninth grader repeated 6th grade
145. P1REPEATG7: Ninth grader repeated 7th grade
146. P1REPEATG8: Ninth grader repeated 8th grade
147. P1REPEATG9: Ninth grader repeated 9th grade
148. P1SLD: Doctor/school has told parent 9th grader has learning disability
149. P1DD: Doctor/school has told parent 9th grader has developmental delay
150. P1AUTISM: Doctor/school has told parent 9th grader has some form of autism
151. P1EAREYE: Doctor/school has told parent 9th grader has hearing/vision problem
152. P1JOINT: Doctor/school has told parent 9th grader has bone/joint/muscle problem
153. P1INTELLECT: Doctor/school has told parent 9th grader has intellectual disability
154. P1ADHD: Doctor/school has told parent 9th grader has ADD or ADHD
155. P1SPECIALED: 9th grader is currently receiving Special Education Services
156. P1LEARN: How much difficulty 9th grader has learning or paying attention
157. P1SPEAK: How much difficulty 9th grader has speaking or communicating
158. P1MOOD: How much difficulty 9th grader has feeling anxious or depressed
159. P1ACTOUT: How much difficulty 9th grader has with behavior problems
160. P1FRIEND: How much difficulty 9th grader has making and keeping friends
161. P1SKIPGRD: Ninth grader has skipped a grade
162. P1CHANGESCH: Number of times 9th grader has changed schools since kindergarten

163. P1SUSPEND: Whether 9th grader has ever been suspended or expelled
164. P1BEHAVE: How often parent contacted by school about problem behavior
165. P1ATTEND: How often parent contacted by school about poor attendance
166. P1PERFORM: How often parent contacted by school about poor performance
167. P1SCHCHOICE: Whether 9th grader's school was assigned or chosen
168. P1SCHMTG: Attended a general school meeting since start of 2009-10 school year
169. P1PTOMTG: Attended a PTO meeting since start of 2009-10 school year
170. P1PTCONFER: Attended parent-teacher conference since start of 2009-10 school year
171. P1SCHEVENT: Attended school event since start of 2009-10 school year
172. P1VOLUNTEER: Served as a school volunteer since start of 2009-10 school year
173. P1FUNDRAISE: Participated in school fund raiser since start of 2009-10 school year
174. P1COUNSELOR: Met with a school counselor since start of 2009-10 school year
175. P1HWOFTEN: How often helped 9th grader with homework
176. P1MTHHWEFF: Confidence in helping with 9th grade math homework
177. P1SCIHWEFF: Confidence in helping with 9th grade science homework
178. P1ENGHWEFF: Confidence in helping with 9th grade English homework
179. P1NOOUTSCH: Didn't participate in any listed out of school activities in last year
180. P1NOACT: Didn't participate in any listed activities with 9th grader in last year
181. P1EDUASPIRE: How far in school would like 9th grader to go
182. P1ABLEBA: 9th grader has ability to complete a Bachelor's degree
183. P1ADMITREQ: Family talked w/ counselor/teacher about postsec admission requirements
184. M1TEACHING: Math teachers in this school set high standards for teaching
185. M1LEARNING: Math teachers in the school set high standards for students' learning
186. M1BELIEVE: Math teachers in this school believe all students can do well
187. M1CLEARGOALS: Math teachers in this school make goals clear to students
188. M1GIVEUP: Math teachers in this school have given up on some students
189. M1CARE: Math teachers in this school care only about smart students
190. M1EXPECT: Math teachers in this school expect very little from students
191. M1WORKHARD: Math teachers in the school work hard to make sure all students learn
192. M1UNPREPPCT: Percentage of students in math course that are unprepared
193. M1ADVBCKGRND: Advanced math course assigned teachers with strongest background
194. M1HELPAVAIL: Rating of availability of Algebra 1 remedial assistance for students
195. M1HELPQUALTY: Rating of quality of Algebra 1 tutoring/remedial assistance for students
196. M1SHRIDEAS: Math teachers in this department share ideas on teaching
197. M1WORKSHOP: Math teachers in dept discuss what was learned at workshop/conference
198. M1SHRSTWRK: Math teachers in this department share and discuss student work
199. M1SHRLESSONS: Math teachers in this dept discuss lessons that were not successful
200. M1SHRBELIEFS: Math teachers in this dept discuss beliefs about teaching/learning
201. M1SHRMTHDS: Math teachers in dept share research on effective teaching methods
202. M1SHRELL: Math teachers in dept share research on ELL instructional practices
203. M1SHRAPPRCH: Math teachers in dept explore approaches for underperforming students
204. M1SHRCONTENT: Math teachers in dept coordinate course content with other teachers
205. M1EFFECTIVE: Math teachers in dept are effective at teaching students in math
206. M1MENTOR: Math teachers in this dept provide support to new math teachers
207. M1CHAIR: Math teachers are supported/encouraged by math department's chair

208. M1TARDY: Student tardiness is a problem at this school
209. M1STUABSENT: Student absenteeism is a problem at this school
210. M1CUT: Student class cutting is a problem at this school
211. M1TCHRABSENT: Teacher absenteeism is a problem at this school
212. M1DROPOUT: Students dropping out is a problem at this school
213. M1APATHY: Student apathy is a problem at this school
214. M1INVOLVEMNT: Lack of parental involvement is a problem at this school
215. M1UNPREPPROB: Students coming unprepared to learn is a problem at this school
216. M1HEALTH: Poor student health is a problem at this school
217. M1RESOURCES: Lack of teacher resources and materials is a problem at this school
218. M1ABLRANGE: Teaching is limited by different academic abilities in the same class
219. M1SESRANGE: Teaching is limited by students with wide range of SES backgrounds
220. M1LANGRANGE: Teaching is limited by students with wide range of language backgrounds
221. M1SPECNEED: Teaching is limited by students with special needs
222. M1UNINTEREST: Teaching is limited by uninterested students
223. M1MORALE: Teaching is limited by low morale among students
224. M1DISRUPT: Teaching is limited by disruptive students
225. M1PROFDEV: Teaching is limited by inadequate professional learning opportunities
226. M1ADMSUPPORT: Teaching is limited by inadequate administrative support
227. M1COMPUTER: Teaching is limited by shortage of computer hardware/software
228. M1TECHSUPPRT: Teaching is limited by shortage of support for using computers
229. M1BOOKS: Teaching is limited by shortage of textbooks for student use
230. M1STUEQUIP: Teaching is limited by shortage of instructional equipment for students
231. M1DEMOEQUIP: Teaching is limited by shortage of equipment for demonstrations
232. M1FACILITIES: Teaching is limited by inadequate physical facilities
233. M1RATIO: Teaching is limited by high student to teacher ratio
234. M1PLANNING: Teaching is limited by lack of planning time
235. M1AUTONOMY: Teaching is limited by lack of autonomy in instructional decisions
236. M1FAMSUPPORT: Teaching is limited by lack of parent/family support
237. M1PRESSURES: School's principal deals w/ outside pressures interfering with teaching
238. M1POORJOBRES: School's principal does poor job of getting resources for this school
239. M1PSETSPRIO: School's principal sets priorities and sees that they are carried out
240. M1PSCHVISION: School's principal communicates kind of school that is wanted to staff
241. M1PCOMEXP: School's principal lets staff members know what is expected of them
242. M1PINNOVATE: School's principal is interested in innovation and new ideas
243. M1PCONSULTS: School's principal consults staff before making decisions affecting them
244. M1TSCHDISC: Teachers at this school help maintain discipline in the entire school
245. M1TIMPROVE: Teachers at this school take responsibility for improving the school
246. M1TSETSTDS: Teachers at this school set high standards for themselves
247. M1TSELFDEV: Teachers at school feel responsible for developing student self-control
248. M1THELPBEST: Teachers at school feel responsible for helping each other do their best
249. M1TALLLEARN: Teachers at this school feel responsible that all students learn
250. M1TFAIL: Teachers at school feel responsible when students in this school fail
251. N1TEACHING: Science teachers in this school set high standards for teaching
252. N1LEARNING: Science teachers in the school set high standards for students' learning

253. N1BELIEVE: Science teachers in this school believe all students can do well
254. N1CLEARGOALS: Science teachers in this school make goals clear to students
255. N1GIVEUP: Science teachers in this school have given up on some students
256. N1CARE: Science teachers in this school care only about smart students
257. N1EXPECT: Science teachers in this school expect very little from students
258. N1WORKHARD: Science teachers in the school work hard to make sure all students learn
259. N1UNPREPPCT: Percentage of students in science course that are unprepared
260. N1ADVBCKGRND: Advanced science course assigned teachers with strongest background
261. N1SHRIDEAS: Science teachers in this department share ideas on teaching
262. N1WORKSHOP: Science teachers in dept discuss what was learned at workshop/conference
263. N1SHRSTWRK: Science teachers in this department share and discuss student work
264. N1SHRLESSONS: Science teachers in this dept discuss lessons that were not successful
265. N1SHRBELIEFS: Science teachers in this dept discuss beliefs about teaching/learning
266. N1SHRMTHDS: Science teachers in dept share research on effective teaching methods
267. N1SHRELL: Science teachers in dept share research on ELL instructional practices
268. N1SHRAPPRCH: Science teachers in dept explore approaches for underperforming students
269. N1SHRCONTENT: Science teachers in dept coordinate course content with other teachers
270. N1EFFECTIVE: Science teachers in dept are effective at teaching students in science
271. N1MENTOR: Science teachers in this dept provide support to new science teachers
272. N1CHAIR: Science teachers are supported/encouraged by science department's chair
273. N1TARDY: Student tardiness is a problem at this school
274. N1STUABSENT: Student absenteeism is a problem at this school
275. N1CUT: Student class cutting is a problem at this school
276. N1TCHRABSENT: Teacher absenteeism is a problem at this school
277. N1DROPOUT: Students dropping out is a problem at this school
278. N1APATHY: Student apathy is a problem at this school
279. N1INVOLVEMNT: Lack of parental involvement is a problem at this school
280. N1UNPREPPROB: Students coming unprepared to learn is a problem at this school
281. N1HEALTH: Poor student health is a problem at this school
282. N1RESOURCES: Lack of teacher resources and materials is a problem at this school
283. N1ABLRANGE: Teaching is limited by different academic abilities in the same class
284. N1SESRANGE: Teaching is limited by students with wide range of SES backgrounds
285. N1LANGRANGE: Teaching limited by students with wide range of language backgrounds
286. N1SPECNEED: Teaching is limited by students with special needs
287. N1UNINTEREST: Teaching is limited by uninterested students
288. N1MORALE: Teaching is limited by low morale among students
289. N1DISRUPT: Teaching is limited by disruptive students
290. N1PROFDEV: Teaching is limited by inadequate professional learning opportunities
291. N1ADMSUPPORT: Teaching is limited by inadequate administrative support
292. N1COMPUTER: Teaching is limited by shortage of computer hardware/software
293. N1TECHSUPPRT: Teaching is limited by shortage of support for using computers
294. N1BOOKS: Teaching is limited by shortage of textbooks for student use
295. N1STUEQUIP: Teaching is limited by shortage of instructional equipment for students

296. N1DEMOEQUIP: Teaching is limited by shortage of equipment for demonstrations
297. N1FACILITIES: Teaching is limited by inadequate physical facilities
298. N1RATIO: Teaching is limited by high student to teacher ratio
299. N1PLANNING: Teaching is limited by lack of planning time
300. N1AUTONOMY: Teaching is limited by lack of autonomy in instructional decisions
301. N1FAMSUPPORT: Teaching is limited by lack of parent/family support
302. N1PRESSURES: School's principal deals w/ outside pressures interfering with teaching
303. N1POORJOBRES: School's principal does poor job of getting resources for this school
304. N1PSETSPRIO: School's principal sets priorities and sees that they are carried out
305. N1PSCHVISION: School's principal communicates kind of school that is wanted to staff
306. N1PCOMEXP: School's principal lets staff members know what is expected of them
307. N1PINNOVATE: School's principal is interested in innovation and new ideas
308. N1PCONSULTS: School's principal consults staff before making decisions affecting them
309. N1TSCHDISC: Teachers at this school help maintain discipline in the entire school
310. N1TIMPROVE: Teachers at this school take responsibility for improving the school
311. N1TSETSTDS: Teachers at this school set high standards for themselves
312. N1TSELFDEV: Teachers at school feel responsible for developing student self-control
313. N1THELPBEST: Teachers at school feel responsible for helping each other do their best
314. N1TALLLEARN: Teachers at this school feel responsible that all students learn
315. N1TFAIL: Teachers at school feel responsible when students in this school fail
316. A1SINGLESEX: Whether school is a single-sex school
317. A1SCHTYPE: School type (charter)
318. A1YRROUND: Whether school is a year round school
319. A1SCHEDULE: Course schedule type
320. A1CLASSHRS: Average instruction hours per day
321. A1ADA: Average daily attendance percentage for high school students
322. A1NOTIFY: Whether parents are notified when students are absent without an excuse
323. A1TRANSFRALT: % of 08-09 students transferred out to an alternative program/school
324. A1G9SUMMER: Offers pre-HS summer reading/math instruction for struggling 9th graders
325. A1G9OVERAGE: Offers learning communities for over-age student lacking HS prerequisite
326. A1G9COMMUNTY: Offers 9th grade learning communities separate from rest of school
327. A1G9DOUBLE: Offers catch-up courses/double-dosing to assist struggling 9th graders
328. A1G9STUDY: Offers study skill seminar/class for struggling 9th graders
329. A1G9TEACHER: Offers assistance for teachers working with struggling 9th graders
330. A1G9TUTOR: Offers tutoring to assist struggling 9th graders
331. A1G9OTHRPROG: Offers another program to assist struggling 9th graders
332. A1CAPACITY: Percent capacity to which school is filled
333. A1OFFERALT: Alternative program offered on-site
334. A1OFFERDOPRV: Dropout prevention program offered on-site
335. A1OFFERAP: College Board Advanced Placement (AP) courses offered on-site
336. A1FREELUNCH: % of student body receiving free or reduced-price lunch
337. A1ELL: % of student body who are English language learners
338. A1SPECIALED: % of student body receiving Special Education services for disabilities
339. A1ALTPROG: % of student body enrolled in an alternative program

340. A1DROPOUTPRV: % of student body enrolled in a dropout prevention program
341. A1AP: % of student body enrolled in Advanced Placement courses
342. A1HISPSTU: % of student body of Hispanic/Latino/Latina origin
343. A1BLACKSTU: % of student body that is Black or African American
344. A1ASIANPISTU: % of student body that is Asian or Pacific Islander
345. A1AMINDIANST: % of student body that is American Indian or Alaska Native
346. A1REPEATG9: % of the 2009-2010 9th-grade class that is repeating 9th grade
347. A1RETURN09: % of 9th graders enrolled in this school Sept 2008 returned Sept 2009
348. A14YRDEGREE: % of 08-09 seniors who went to 4-year Bachelor's-granting institution
349. A12YRDEGREE: % of 08-09 seniors who went to Associates-granting/technical institution
350. A1WORK: % of 08-09 seniors who entered the workforce
351. A1MILITARY: % of 08-09 seniors who joined military
352. A1FTTCHRS: Total number of full-time teachers
353. A1PTTCHRS: Total number of part-time teachers
354. A1IB: School offers an International Baccalaureate (IB) program
355. A1SEX: Principal's sex
356. A1HISP: Principal is of Hispanic/Latino/Latina origin
357. A1BLACK: Principal is Black or African American
358. A1ASIAN: Principal is Asian
359. A1AMINDIAN: Principal is American Indian/Alaska Native
360. A1HIDEG: Principal's highest degree earned
361. A1MANAGEMENT: Prior management experience outside of the field of education
362. A1ALTPREP: Whether became a principal through alternative prep program
363. A1CERTIFIED: Principal is certified as a principal in this state
364. A1YRSADMIN: Years served as principal of any school
365. A1YRSHSLSSCH: Years served as principal of this school
366. A1TEACHING: Principal is currently teaching in this school
367. A1YRSMSTCHR: Principal's years of middle school teaching experience
368. A1YRSHSTCHR: Principal's years of secondary teaching experience
369. A1HRTEACHERS: Hours/week spent working with teachers on instructional issues
370. A1HRINTMGMNT: Hours/week spent on internal school management
371. A1HREXTMGMNT: Hours/week spent on external school management
372. A1HRDISCIPLN: Hours/week spent on student discipline/attendance
373. A1HRMONITOR: Hours/week spent monitoring hallways/campus/lunchroom
374. A1HRTEACHING: Hours/week spent on principal's own teaching assignments
375. A1HRPARENT: Hours/week spent talking and meeting with parents
376. A1HRSTUDENT: Hours/week spent meeting with students
377. A1HRPAPERWK: Hours/week spent on paperwork required by authorities
378. A1HROTH: Hours/week spent on other activities
379. A1TARDY: Student tardiness is a problem at this school
380. A1STUABSENT: Student absenteeism is a problem at this school
381. A1CUT: Student class cutting is a problem at this school
382. A1TCHRABSENT: Teacher absenteeism is a problem at this school
383. A1DROPOUT: Students dropping out is a problem at this school
384. A1APATHY: Student apathy is a problem at this school
385. A1PRNTINV: Lack of parental involvement is a problem at this school

386. A1UNPREP: Students coming unprepared to learn is a problem at this school
387. A1HEALTH: Poor student health is a problem at this school
388. A1RESOURCES: Lack of teacher resources and materials is a problem at this school
389. A1CONFLICT: Frequency of physical conflicts among students at this school
390. A1ROBBERY: Frequency of robbery or theft at this school
391. A1VANDALISM: Frequency of vandalism at this school
392. A1DRUGUSE: Frequency of student illegal drug use at this school
393. A1ALCOHOL: Frequency of students use of alcohol while at school
394. A1DRUGSALE: Frequency of drug sales on the way to/from school or on school grounds
395. A1WEAPONS: Frequency of student possession of weapons at this school
396. A1PHYSABUSE: Frequency of physical abuse of teachers at this school
397. A1TENSION: Frequency of student racial tensions at this school
398. A1BULLY: Frequency of student bullying at this school
399. A1VERBAL: Frequency of student verbal abuse of teachers at this school
400. A1MISBEHAVE: Frequency of student in-class misbehavior at this school
401. A1DISRESPECT: Frequency of student acts of disrespect for teachers at this school
402. A1GANG: Frequency of student gang activities at this school
403. C1FTCNSL: Number of full-time high school counselors
404. C1PTCNSL: Number of part-time high school counselors
405. C1FTCERTCNSL: Number of certified full-time high school counselors
406. C1PTCERTCNSL: Number of certified part-time high school counselors
407. C1CASELOAD: Average caseload for school's counselors
408. C1HRSSCHED: % hours counseling staff spent on high school course choice/scheduling
409. C1HRSCOLLEGE: % hours counseling staff spent on college readiness/selection/apply
410. C1HRSCAREER: % hours counseling staff spent on occupational choice/career planning
411. C1HRSDEVELOP: % hours counseling staff spent on personal/academic/career development
412. C1HRSJOBSKLL: % hours counseling staff spent on job placement/job skill development
413. C1HRSPROBLEM: % hours counseling staff spent on school/personal problems
414. C1HRSTESTING: % hours counseling staff spent on academic testing
415. C1HRSNONCNSL: % hours counseling staff spent on non-counseling activities
416. C1HRSOTHCNSL: % hours counseling staff spent on other counseling activities
417. C1GOAL1: School counseling program's most emphasized goal
418. C1GOAL2: School counseling program's second most emphasized goal
419. C1GOAL3: School counseling program's third most emphasized goal
420. C1TRANSCNSL: MS counselors meet with HS counselors to assist with student transition
421. C1TRANSCRS: HS counselors meet with 8th graders to select 9th grade courses
422. C1TRANPRNT: HS counselors present HS course/registration information to MS parents
423. C1TRANPRES: HS counselors present HS course/registration information to MS students
424. C1TRANCOTH: HS counselors assist students with transition from MS to HS in other way
425. C1TRANNOT: HS counselors do not assist students with transition from MS to HS
426. C1TRANSTUDPR: HS students present information at MS to assist with student transition
427. C1TRANSTFFPR: HS staff present information at MS to assist with student transition
428. C1TRANVISIT: Before school year MS students are invited to HS social event
429. C1TRANCLASS: MS students attend regular classes at HS
430. C1TRANADMIN: MS and HS administrators meet together on articulation and programs

431. C1TRANTCHRS: MS and HS teachers meet together on courses and requirements
432. C1TRANBUDDY: Buddy or big brother/sister programs pair new students with older ones
433. C1TRANLRNCOM: 9th graders are placed in small learning communities/9th Grade Academies
434. C1TRANSUMMER: Parents/students visit the HS during summer before students enter HS
435. C1TRANFALL: Parents visit HS for orientation in fall after children have entered
436. C1TRANSOTH: School assists with transition from MS to HS in some other way
437. C1TRANNONE: School offers no assistance to students transitioning from MS to HS
438. C1STRUGGLE: School offers summer enrichment courses to struggling students
439. C1TUTOR: Tutoring during school day is available for students needing extra help
440. C1STAFF: Staff work with teachers to provide extra help for students
441. C1PULLOUT: Pull-out instruction during school day for students needing extra help
442. C1CREDREC: Off-track/day/evening/summer school credit recovery program is available
443. C1HOMEWORK: Homework assistance program available for students needing extra help
444. C1OUTSIDE: Support outside the school day for students needing extra help
445. C1OTHRASSIST: School takes other steps to assist struggling high school students
446. C1DROPOUT: School has a formal dropout prevention program for high school students
447. C1GEDPREP: School has formal GED test preparation program on-site
448. C1CTE: CTE or vocational-technical program offered
449. C1TTEACHING: Teachers in this school set high standards for teaching
450. C1TLEARNING: Teachers in this school set high standards for students' learning
451. C1TBELIEVE: Teachers in this school believe all students can do well
452. C1TGIVEUP: Teachers in this school have given up on some students
453. C1TCARE: Teachers in this school care only about smart students
454. C1TEXPECT: Teachers in this school expect very little from students
455. C1TWORKHARD: Teachers in this school work hard to make sure all students learn
456. C1CLEARNING: Counselors in this school set high standards for students' learning
457. C1CBELIEVE: Counselors in this school believe all students can do well
458. C1CGIVEUP: Counselors in this school have given up on some students
459. C1CCARE: Counselors in this school care only about smart students
460. C1CEXPECT: Counselors in this school expect very little from students
461. C1CWORKHARD: Counselors in this school work hard to make sure all students learn
462. C1PLEARNING: Principal in this school sets high standards for students' learning
463. C1PBELIEVE: Principal in this school believes all students can do well
464. C1PGIVEUP: Principal in this school has given up on some students
465. C1PCARE: Principal in this school cares only about smart students
466. C1PEXPECT: Principal in this school expects very little from students
467. C1PWORKHARD: Principal in this school works hard to make sure all students learn
468. C1ENCCLG: School has program to encourage student not considering college to do so
469. C1CLGPREP: School has counselor designated for college readiness/selection/apply
470. C1WORKFORCE: School has counselor designated for workforce preparation/placement
471. C1CLGFAIR: School holds or participates in college fairs
472. C1POSTSECREQ: School consults with postsecondary representatives about requirement/qualifications
473. C1VISITCLG: School organizes student visits to colleges

474. C1UPBOUND: School offers college preparation programs - Upward Bound/GEAR UP/AVID/MESA
475. C1INFOSESSN: School holds info session on transition to college for students/parents
476. C1FINANCEAID: School assists students with finding financial aid for college
477. C1DUALENROLL: School provides opportunities for dual/concurrent enrollment
478. C1BEHAVIOR: School offers counseling curriculum for positive academic behaviors
479. C1ASSISTOTH: School takes other steps to assist with HS to college transition
480. C1INTERN: School offers internships with local employers
481. C1JOBFAIR: School offers job fairs
482. C1JOBGUIDE: School offers career guides or skills assessments
483. C1EMPLOYER: School offers school/classroom presentations by local employers
484. C1AWARENESS: School offers career awareness activities
485. C1DECISION: School offers courses in career decision making
486. C1WORKSTUDY: School offers exploratory work experience programs/co-op/work-study/EBCE
487. C1CAREERDAY: School offers career days or nights
488. C1ASSEMBLIES: School offers vocational oriented assemblies and speakers in classes
489. C1VOCTECH: School offers vocational-technical courses not part of formal program
490. C1JOBVISIT: School offers job site visits/field trips
491. C1JOBSHADOW: School offers job shadowing
492. C1JOBTEST: School offers tests for career planning purposes
493. C1JOBSKILLS: School offers training in job seeking skills
494. C1JOBINFOCMP: School offers computerized career information resources
495. C1JOBINFONON: School offers non-computerized career information resources
496. C1HSTOWORKNO: School doesn't assist students with transition from high school to work

A.2 Technical Details

A.2.1 Supervised ML: Implementation Details and Robustness Checks

We have implemented LASSO in Stata using the user-written code by Christian Hansen[19]. We have included all possible predictors as inputs in the algorithms. We have then used for the LASSO with all these variables a grid-search among 22 values (plus the default one[20]) to find the optimal penalization term, i.e. the one that maximizes the recall rate subject to a minimum accuracy rate. In other words, we have used LASSO with a certain penalization term to select the top predictors in the training sample, used the selected variables as inputs in an OLS or Logit model, selected the penalization term which maximizes the post-LASSO performances in the CV sample, and recorded the out-of-sample performances. Table 2 reports the 5-fold averages obtained repeating the above procedure for each different combination of train, CV and test sample.

The post-LASSO with interaction terms (Table 2 Model 5) has been obtained by running LASSO twice. First, LASSO has been used to select the top 25 predictors among all the variables available. Second, we have generated all possible two-way interactions among these variables, as well as quadratic and cubic functions of the non-binary selected variables. Third, we have used all these interaction and higher order terms, together with the whole set of variables available, as inputs in the LASSO model. The overall number of inputs was around 2,050. We have then selected the optimal penalization term using a grid search and reported the post-LASSO performances as described in the previous paragraph.

The post-LASSO with school fixed effect model (Table 2 Model 6) has been obtained by using LASSO to select the top predictors among the individual variables derived from the student and parent questionnaire, and by then including these variables, together with the school fixed effect, in a Logit model. We have reduced the minimum accuracy rate from 89% to 87% because none of the penalization terms considered provided such a sufficiently high accuracy.

It is also important to point out that, in cases in which a variable perfectly predict dropout for some observations, Stata drops both such observations and variable when running the Logit model. This has led to slightly lower number of observations in Table 2 Models 4-5 (less than 50 observations dropped), while there are much more substantial reductions when including school FE in Table 2 Model 6 (the test sample size goes from 4,290 to around 3,000). However, SVM, Boosting and OLS Post-LASSO have the same number of observations as the basic models (Table 1) and still higher performances, thus reassuring us that our conclusions are not driven by changes in the sample size.

We have implemented the SVM algorithm in Stata using the command *svmachines* (Guenther and Schonlau, 2016). In order to reduce the computational time, we have used LASSO to select around 350 predictors as inputs variables in the SVM algorithm. A grid-search has been used to find the optimal parameters, i.e. the ones that maximize the recall rate subject to a minimum accuracy rate. We have considered 6 possible values for each parameter (the penalization term

[19] The code can be downloaded from http://faculty.chicagobooth.edu/christian.hansen/research/#Code.

[20] Although it is more appropriate when dealing with model selection, not prediction, we have also included the default value for the penalization term among the possible candidates in the grid-search. As discussed in Belloni et al. (2014), such default value depends on the numbers of predictors and observations included in the LASSO algorithm. Its theoretical justification and derivation is provided in Belloni et al. (2012) and Belloni et al. (2013).

and the kernel smoothing parameter), as well as two kernels (Guassian and Sigmoid) for a total of 72 combinations. SVM does not produce estimated probabilities: it only predicts on which side of the margin is an observation. Therefore, when computing the AUC, we have estimated pseudo-probabilities through Platt Scaling (Guenther and Schonlau, 2016). However, in this case data are sorted randomly using the operating system random-number generator. As a consequence, different runs may give different results, potentially making the result for the AUC-SVM in Table 2 not exactly replicable in other computers.

Boosted regression has been implemented in Stata using the command *boost* (Schonlau, 2005). A grid-search has been used to find the optimal parameters. We have considered two values for the shrinkage parameter, two error distribution (Gaussian and Logistic), up to 40 iterations (i.e. number of trees), and each tree could have up to 5 splits, for a total of 800 possible combinations (5*40*2*2). Introducing a shrinking parameter reduces the contribution of each additional tree, thus decreasing the impact of an over-fitted tree. The cost of this procedure is a substantial increase in computational time, since it is necessary to increase the number of iterations in order to compensate these smaller steps. Bagging is an additional technique which we have used to reduce the variance of the final prediction without influencing the bias. At each iteration, we have used only a random subset of the train set (80% of it) to build the tree.

It is worth mentioning that the calibration procedure has succeed in avoiding over-fitting the data. For instance, the recall rates in the test samples over the five folds considered varied between 17.6% and 24.1%, while such rates were between 22.1% and 33.5% in the train samples (20.0% - 24.7% in the CV sample). Therefore, there is not a big gap between the performances in these sets of samples, implying that the algorithm is not suffering from high variance.

As discussed in Section 2.2, the dependent variable has been set equal to one if the student, school or parent had reported at least one known dropout episode in one of the interviews. By definition, if such information was not available, e.g. if the student did not reply in the last follow-up, the student was not counted as dropout. Excluding non-respondents and students whose status was unknown from the estimation does actually improve the recall rate (even if it reduces the sample size to around 16,400 observations). For instance, if we replicate the same Logit Post-LASSO algorithm as Model 4 in Table 2 for this alternative outcome variable, we obtain an AUC of 0.83, an accuracy rate of 89.3%, and a recall rate of 35.7%.

A.2.2 Microeconomic Foundation: Technical Details

This Appendix contains the detailed explanation of how the results in Section 3.6 have been obtained.

Figure 1 has been plotted using the Stata built-in command *roctab*. We have estimated five logit models, one for each training set in the 5-fold CV procedure. For each specification, after having estimated it using the relevant training sample, we have computed the predicted probabilities for each observation in the corresponding test sample. Since there are no intersections between the five test samples (each of them comprises 20% of the data), we thus have one predicted probability for each observation. The ROC curve is then constructed by varying between 0 and 1 the threshold used to translate predicted probabilities into predicted outcomes (binary variable 0/1). In particular, each realized value among the predicted probabilities is used as possible threshold. Finally, predicted outcomes are compared with actual outcomes in order to compute sensitivity and sensibility given any possible threshold. Note that the area under the curve is

slightly different than the one reported in Table 1 (0.79 vs. 0.80): this is due to the fact that the former is computed after combining the predicted probabilities in the five folds, while the latter was computed as the average of the area under the five different ROC curves, one for each fold.

In order to construct Table 6, for each fold we have estimated a logit model using the corresponding training sample. We have then computed the predicted probabilities and selected the threshold which maximizes the recall rate in the CV sample while satisfying the budget constraint. Such optimal threshold has then been used to compute the predicted outcomes in the corresponding test sample. For each fold, we have then stored for the test sample the accuracy, recall rate and actual cost of the intervention. Table 6 reports the 5-fold averages for these variables, as well as the average of the five optimal thresholds. We have then repeated the simulation for each possible combination of cost per student and overall budget.

A.2.3 Unsupervised ML: Technical Details

As discussed in Hastie et al. (2009) and Stata (2015), there are several methods to divide observations into groups based on some variables $\boldsymbol{X}$. All these methods are referred to as unsupervised ML algorithms since there is no observable dependent variable: researchers need to understand whether individuals can be partitioned into different clusters only based on some of their characteristics. More formally, cluster analysis specifies whether the joint density of $\boldsymbol{X}$, $P(\boldsymbol{X})$, can be represented by a mixture of simpler densities representing distinct groups of observations.

One of the most common unsupervised ML is hierarchical clustering. Conceptually, the hierarchical clustering algorithm can be summarized as follows. Initially there are n distinct groups, one for each observation. In the next step, the two closest observations are merged into one group, thus resulting in $n-1$ groups. After that, the closest two groups are merged together, producing $n-2$ groups. This process continues until all the observations are merged into one large group. Therefore, the output of this algorithm is a hierarchy of groupings from one group to n groups.

There are four decisions involved in this procedure: measuring distance between observations, measuring distance between groups, selecting the number of observable variables, and selecting the optimal number of groups.

In order to measure the distance between observations, we have used the traditional Euclidean distance, i.e. a Minkowski distance metric with argument 2. Give two observations x_i and x_j with p variables, this can be defined as:

$$\left[\sum_{h=1}^{p}\left(x_{ih}-x_{jh}\right)^2\right]^{\frac{1}{2}}$$

The distance between two groups can be measured by considering the two closest observations between them (single linkage), the farthest observations between them (complete linkage), their means (centroids), or the average (dis)similarity between the observations of the two groups (average linkage). There are additional definitions of distance available (e.g. median linkage, Ward's linkage, and weighted average linkage), but we have selected average linkage since its performances are good in several simulations, thus it is reasonably robust (Stata, 2015).

As far as variable selection is concerned, there are some procedures available to mechanically select the number of variables which are then used by the clustering algorithm to group observations together (Witten and Tibshirani, 2010). Nevertheless, we have used the same variables selected by the Logit Post-LASSO algorithm at least 3 times (Table 4). We believe that this approach is more coherent with the previous sections and less subject to arbitrary variable selections.

Finally, although there is no clear-cut solution to this issue, in order to select the optimal number of groups we have used two stopping rules: the Caliński and Harabasz pseudo-F index and the Duda-Hart *Je(2)/Je(1)* index with the associated pseudo-T^2. Distinct clustering is signaled by a high Caliński and Harabasz pseudo-F index, as well as by a large *Je(2)/Je(1)* index associated with a low pseudo-T^2 surrounded by much larger pseudo-T^2 values. The general idea behind these stopping rules is to ask whether splitting one cluster would reduce a certain loss function or another measure of fit. The Caliński and Harabasz pseudo-F index compares the sum of squared distances within the partitions - that is, the distances between clusters - to that in the unpartitioned data, taking account of the number of clusters and number of cases. Formally, if we have q groups ($C_1 \dots C_q$) and n observations:

$$pseudo - F_{CH} = \frac{trace(B_q)/(q-1)}{trace(W_q)/(n-q)}$$

$$B_q = \sum_{k=1}^{q} |C_k| \|\bar{c}_k - \bar{x}\|^2$$

$$|C_k| = \sum_{i=1}^{n} \mathbb{1}\{x_i \in C_k\}$$

$$W_q = \sum_{k=1}^{q} \sum_{i=1}^{n} \mathbb{1}\{x_i \in C_k\} \|x_i - \bar{c}_k\|^2$$

Where $\bar{x}$ is the centroid of the data, $\bar{c}_k$ is the centroid of the generic cluster C_k, and x_i is the vector of characteristics for individual i. In other words, $|C_k|$ is the number of elements in cluster C_k, B_q is the between-group dispersion matrix for the data clustered into q clusters (the weighted sum of distances between the group centroids and the data centroid), and W_q is the within-group dispersion matrix for the data clustered into q clusters (the sum of the distance of each observation from its groups centroid).

The Duda-Hart *Je(2)/Je(1)* index is literally *Je(2)*, the sum of squared errors within clusters in the two derived clusters (C_h and C_l), divided by *Je(1)*, the sum of squared errors in the combined original cluster (C_m).

$$Duda - Hart = \frac{Je(2)}{Je(1)} = \frac{W_h + W_l}{W_m}$$

Where W is defined as in the Caliński and Harabasz pseudo-F index above. In addition to this, the Duda-Hart pseudo-T^2 statistic takes account of the number of observations in both clusters (n_h and n_l):

$$\frac{1}{Je(2)/Je(1)} = 1 + \frac{T^2}{n_h + n_l - 2}$$

Using these criteria, we have decided to divide our sample in 10 groups. After that, we have excluded the groups with only few observations since it was not possible to obtain meaningful summary statistics with such small sample sizes.

A.3 Additional Tables

Table A1: Variables selected by Boosting at least 2 times (out of 5)

Predictors	Count	Influence
GPA in 9th grade	5	39.7
Born in 1993 (most students were born in 1994-1995)	5	11.2
HSLS:09 Math test score	4	5.9
Whether 9th grader has ever been suspended or expelled	4	5.2
GPA for all academic 9th grade courses	4	2.5
Parent contacted by school about poor attendance more than 4 times	4	2.4
Born in 1992	4	1.9
No science courses taken in 9th grade	3	10.8
No math courses taken in 9th grade	3	4.1
9th grader very sure that he/she will graduate from high school	3	1.5
Credits earned in 9th grade	3	1.3
Number of household members	3	1.1
9th graders has changed schools 7 times since kindergarten	3	0.4
GPA in 9th grade missing	2	6.3
Born in 1995	2	4.5
Parent participated in school fundraiser	2	1.7
Parent thinks 9th grader will at most attain HS	2	1.4
Born in 1994	2	1.4
9th grader spend less than 1h/day on extracurricular activities	2	1.3
% student body receiving free lunch	2	1.2
9th graders has never changed schools since kindergarten	2	1.1
Parent reporting a little difficulty by 9th grader with behavior problems	2	1.0
% student body enrolled in dropout prevention program	2	0.8
9th graders has changed schools 6 times since kindergarten	2	0.8
% 08-09 transferred out to an alternative program/school	2	0.7
% student body receiving special education services for disabilities	2	0.7
Principals' years of experience in the school	2	0.7
% 08-09 seniors who entered the workforce	2	0.7
% 08-09 seniors who went to Associates-granting/technical institutions	2	0.6
% student body of Hispanic origin	2	0.6
% student body that is American Indian or Alaska Native	2	0.4
9th grader was homeschooled in previous academic year	2	0.4
9th grader thinks that even if he/she studies, he/she won't get into college	2	0.3

References cited only in the Online Appendix

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. Econometrica 80, 2369–2429. doi:10.3982/ECTA9626

Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. J. Econ. Perspect. 28, 29–50. doi:10.1257/jep.28.2.29

Belloni, A., Chernozhukov, V., Hansen, C., 2013. Inference on treatment effects after selection among high-dimensional controls. Rev. Econ. Stud. 81, 608–650. doi:10.1093/restud/rdt044

Guenther, N., Schonlau, M., 2016. Support Vector Machines. Stata J. 16, 917–937.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edi. ed, Springer Series in Statistics. Springer. doi:10.1007/b94608

Ingels, S.J., Pratt, D.J., Herget, D.R., Burns, L.J., Dever, J.A., Ottem, R., Rogers, J.E., Jin, Y., Leinwand, S., 2011. High School Longitudinal Study of 2009 (HSLS:09). Base-Year Data File Documentation. Washington, DC.

Ingels, S.J., Pratt, D.J., Herget, D.R., Dever, J.A., Fritch, L.B., Ottem, R., Rogers, J.E., Kitmitto, S., Leinwand, S., 2014. High School Longitudinal Study of 2009 (HSLS:09) Base Year to First Follow-Up Data File Documentation. Washington, DC.

Schonlau, M., 2005. Boosted regression (boosting): An introductory tutorial and a Stata plugin. Stata J. 5, 330–354. doi:The Stata Journal

Stata, 2015. Stata Base Reference Manual [MV], Release 14. ed. Stata Press.

Witten, D.M., Tibshirani, R., 2010. A Framework for Feature Selection in Clustering. J. Am. Stat. Assoc. 105, 713–726.