OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 81, 2 (2019) 0305–9049
doi: 10.1111/obes.12277

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

DARIO SANSONE†

†*Department of Economics, Georgetown University, ICC 580, 37th and O Streets, N.W., Washington, DC 20057-1036, USA (e-mail: ds1289@georgetown.edu)*

Abstract

This paper combines machine learning with economic theory in order to analyse high school dropout. It provides an algorithm to predict which students are going to drop out of high school by relying only on information from 9th grade. This analysis emphasizes 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. Goodness-of-fit criteria are selected based on the-context and the underlying theoretical framework: model parameters are calibrated by taking into account the policy goal – minimizing the expected dropout rate - and the school budget constraint. Finally, this study verifies the existence of heterogeneity through unsupervised machine learning by dividing students at risk of dropping out into different clusters.

I. Introduction

High school dropout is a key issue in the US educational system: only 83.2% of students graduated with a regular high school diploma within 4 years of starting 9th grade in 2015. According to the OECD (2016), the US upper-secondary graduation rate of 82% is below average among advanced economies (85%), and far from the graduation rates in Germany (91%), Japan (97%) and Finland (97%). Furthermore, there are substantial gender, racial and geographical gaps within the United States (IES, 2016).[1]

This issue has been extensively analysed by researchers in economics and public policy (De Witte *et al.*, 2013; Murnane, 2013). The U.S. Department of Education provided

JEL Classification numbers: C53; C55; I20.

*I am grateful to the Editor Brian Bell, one anonymous referee, Garance Genicot, Francis Vella, Laurent Bouton, Daniel Ackerberg, Pooya Almasi, Mary Ann Bronson, Nick Buchholz, Benjamin Connault, Francis DiTraglia, Luca Flabbi, Myrto Kalouptsidi, Madhulika Khanna, Ivana Komunjer, Gizem Kosar, Arik Levinson, Whitney Newey, Hiren Nisar, Elena Arias Ortiz, Franco Peracchi, Mariacristina Rossi, John Rust, Bernard Salanié, Shuyang Sheng, Arthur van Soest, Allison Stashko, Basit Zafar and participants to the 2018 SOLE Conference, 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] It should be mentioned 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).

almost $1.5 billion in grants to schools investing in innovative practices aimed at increasing graduation rates between 2010 and 2016 (Office of Innovation & Improvement, 2016). Failing to graduate from high school has high costs, as only 12% of all jobs in the economy will require less than a high school diploma by 2020 (Carnevale, Smith and Strohl, 2013). Schooling also has several non-pecuniary benefits ranging from health to happiness, marriage, trust, and work enjoyment (Oreopoulos, 2007; Oreopoulos and Salvanes, 2011).

This paper shows how machine learning (ML) and economic theory can be jointly applied in education. In particular, this paper creates a model that identifies students who are at risk of dropping out using information from their first year of high school. In doing so, it also illustrates how ML can be used to identify top predictors and heterogeneity among students. In addition, the first part of this paper demonstrates that trying to predict vulnerable students using a limited number of educational variables can detect only a small fraction of those students who actually end up dropping out of high school. This result is especially relevant since 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 behaviour and course grades to find students at-risk, even when there is minimal empirical evidence to support this recommendation (Rumberger *et al.*, 2017). In contrast to these practices, this paper shows how schools can exploit available big data, jointly with ML techniques, to substantially improve these predictions. These more advanced algorithms have the potential to correctly identify thousands of additional students who are at risk of dropping out every year.

After having identified vulnerable students, this paper illustrates the application of unsupervised ML to cluster such individuals into different groups based on their observable characteristics. Clustering students has two advantages. First, it emphasizes that these students are not a homogeneous group: 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 neighbourhoods or they come from very poor households. The latter group would likely require different programmes than the first one. Tutoring might be more appropriate for students struggling in certain subjects, while combining tutoring with counselling might be more effective for students with disadvantaged backgrounds. ML can therefore be used to identify students at-risk, and to help design treatments appropriate for each sub-population. Second, it is possible to evaluate how a policy has different impacts among students in various clusters. Indeed, any dropout prevention programme can have different effects depending on student's gender, race, ability, income, as well as by sub-populations. In this way, it is possible to 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 since its goal is to maximize out-of-sample prediction. Algorithms can identify patterns too subtle to be detected by human observations (Luca, Kleinberg and Mullainathan, 2016), thus outperforming econometric models built using heuristic or theory-based approaches. Although there are several policy-relevant issues that do not require causal inference, but rather accurate predictions (Kleinberg *et al.*, 2015), ML applications have been quite limited in economics so far. However, ML is gaining momentum (Belloni, Chernozhukov

and Hansen, 2014; Varian, 2014; McKenzie and Sansone, 2017; 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).

A disadvantage of using off-the-shelf ML techniques to tackle classification problems – applications where the dependent variable is discrete – is that there is no unique method to measure performance. Practitioners generally adopt rules-of-thumb and criteria such as pseudo-R^2 and accuracy (Bowers, Sprott and Taff, 2013), but they often do not justify the reason behind such choices. This paper builds 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, as also advocated in Subrahmanian and Kumar (2017). In other words, a school's constrained optimization problem is taken into account while calibrating the algorithms to maximize prediction performances. Therefore, this paper provides a microeconomic foundation to the choice of the particular criterion used in the paper to select the optimal values of the model parameters and to evaluate the algorithms.

Despite the aforementioned limitation, ML approaches provide several advantages. First, they offer an inexpensive alternative to the numerous tests and assessments that are used to sort and categorize students since kindergarten (Shields, Cook and Greller, 2016). Second, since these algorithms use only information from 9_{th} grade, school counsellors and teachers can detect students at-risk before it is too late to intervene. Even if some scholars have argued for a focus on early childhood education (Carniero and Heckman, 2003), recent studies have shown that it is possible to design effective interventions for disadvantaged and low-skilled adolescents (Cook *et al.*, 2014; Fryer, 2017), thus supporting the need to identify the students most at risk of dropping out. For instance, Cortes, Goodman and Nomi (2015) found that double-dose algebra in 9th grade targeted towards below-average math students increased high school graduation. Similarly, Rodriguez-Planas (2012) estimated that low-performing 9th graders were more likely to graduate on-time when assigned to a programme offering mentoring, educational services, and financial rewards. At the same time, researchers have reported limited or even negative effects of universal programmes that require all students to take college preparatory courses (Allensworth *et al.*, 2009; Clotfelter, Ladd and Vigdor, 2015), thus emphasizing the importance of targeting intervention in order to improve their efficiency and impacts.

To summarize, this paper applies modern state-of-the-art techniques to improve schools' response to elevated school dropout rates. In addition to this, it combines economic theory with ML to adapt these tools to the specific educational context. Finally, it introduces unsupervised ML as a first step toward offering more personalized treatments to students at risk of dropping out.

II. Data

Data source

This paper uses the High School Longitudinal Study of 2009 (HSLS:09), which is a panel micro study interviewing around 21,440 students in 9th grade from about 940 participating schools. The survey design has two levels. First, private and public schools were selected

at the national level. Second, around 30 students were randomly drawn among 9th graders from every school selected in the previous step.

In the first round, information was collected from the selected 9th graders, their parents, math and science teachers, school administrators and lead school counsellors. 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, and a brief update was conducted in 2013 (summer and fall) to record students' postsecondary plans. Students, parent, school administrators and counsellors were interviewed again in 2012. This wave did not include new questionnaires for teachers. Finally, only students and parents were interviewed in 2013.

A math assessment was first administered to students in 9th grade (2009), and then in 11th grade (2012). Data from the students' transcripts including their GPA, their AP class grades, their SAT scores, and the number of credits taken in each subject during high school are also available.[2]

From a policy perspective, the use of the HSLS:09 implies another substantial contribution of this paper. The results presented in the empirical analysis not only focus on the general issue of high school dropout, but are derived from data on 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 NLSY:79, which are attractive since they contain a rich variety of information and span over several decades, but they estimate parameters which may have changed over time, thus lacking external validity.

Outcome variable

The aim of section III is to predict who is eventually going to drop out of high school using information available in 9th grade, i.e. in the first year of high school. Notably, 45% of the schools in the sample had a formal dropout prevention programme in 2009. These programmes included a variety of initiatives: the most common were tutoring and graduation counselling, but some schools also offered job counselling, childcare for students' children, occupational-focused courses, or even incentives for better attendances and classroom performance. When school counsellors were asked in the HSLS:09 how students were selected in order to participate in these programmes, the two most common answers indicate a focus on individuals with poor grades (93%) and fewer 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 known dropout episode regarding the student, and zero otherwise. It is important to note that alternative completers (such as GED recipients) are considered as dropouts. This is in line with the literature that emphasizes the differences between GED recipients and high school graduates (Heckman and Rubinstein, 2001; Heckman, Humphries and Mader, 2011; Zajacova, 2012). Non-respondents are counted as zero. Re-taking a year is also not considered equivalent to dropping out of school.

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

Among the interviewed students, almost 11% had at least one known dropout episode before the second follow-up interview. It is important to note that 82% of the schools in the sample had at least one interviewed student with a recorded dropout episode. In line with the findings from other studies (Adelman *et al.*, 2018), dropouts were not concentrated in a few schools. Therefore, merely targeting low-performing schools would lead to substantial misallocation of resources.

III. Predictions

Technical considerations

Before showing the results from the prediction analysis, it is important to highlight a few technical points. The first one concerns over-fitting, i.e. having a high in-sample predictive power, but a 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 performances in the cross-validation sample (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 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. The in-sample and out-of-sample performances are estimated five times – one for each data split – and the 5-fold average out-of-sample performances are then reported. The k-fold CV is a rather common resampling technique, and while there is no formal rule, 5 or 10 is the usual choice for k since it is computationally less burdensome than other techniques such as the leave-one-out cross-validation and it performs well in simulations (Kuhn and Johnson, 2013).

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 that contains only a constant. The second class comprises all the indices that 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 is typically also interpreted as follows:

		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:

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

However, when classes are imbalanced as in this application, 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, it might be desirable to select a model with lower accuracy but higher predictive power; that is, a model performing better under alternative performance metrics. The criteria which are commonly used are:

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

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

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

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 or rules of thumb. In what follows, the analysis focuses on the recall rate since 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. Section Microeconomic foundation formally justifies this choice using a microeconomic constrained optimization model.

Finally, almost all algorithms (with the notable exception of Support Vector Machines) produce predicted probabilities. The models in section A Basic model follow the convention to predict one when such probability is equal or <0.5, zero otherwise. This is in line with the Bayes classifier (Hastie, Tibshirani and Friedman, 2009), where accuracy rate is maximized by assigning each observation to the most likely class, given its predicted

TABLE 1

Basic model (5-fold average)

	Algorithm	*Inputs*			*Performances*		
		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: This table reports out-of-sample performances of different models estimating the probability that a student drops out of high school. *Individual* indicates that the algorithm uses 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.

probabilities. Lower thresholds lead to higher recall rates, but lower accuracy. Subsequently, section Microeconomic foundation shows how to choose such cut-off during the CV procedure in order to optimize the school objective function. It also illustrates 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 commonly used as a performance criterion.

A basic model

As discussed in section outcome variable, most schools select students who need to participate in dropout prevention programmes based on their past achievement (GPA and number of credits). Therefore, a natural way to start the analysis is to test the power of these predictors. In other words, it is possible to estimate a simple logit model using as regressors student past performances, school attendance and behaviour, as well as all the others variables highlighted in the literature: demographics, school characteristics, and family background.[3] As shown in Table 1 Model 1, the performances are strikingly low: even though the average out-of-sample accuracy rate is almost 90%, the recall rate is just 15%. This means that only a small percentage of the students in the test sample who did eventually end up dropping out are identified as at-risk. These performances are even worse for the OLS and Probit estimates (Model 2 and 3 respectively).

These results do not depend on the sample size. Similar accuracy and recall rates are also obtained when random subsets of the train sample are used (e.g. 30%, 50%, 80%). The average in-sample accuracy and recall rates for the Logit model are around 90% and 15.7%,

[3] In particular, the following 28 variables have been selected: 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, 5-fold CV procedure has also been implemented in these simple models. Moreover, in order to maintain the same number of observations across specifications, missing values have been imputed to zero while adding an indicator variable for such 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.

thus close to the out-of-sample performances. Moreover, even when the Logit model is estimated using all the available observations instead of the 60% training set, the in-sample recall rate is only 17% (92% accuracy, 0.81 AUC). Therefore, collecting data on additional students would not improve these predictions: the algorithm is suffering from high bias (under-fitting). Including more training observations would not solve this issue.

Interactions terms can be added to take into account potential heterogeneity and use 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, including 14 interaction terms does not improve performances.

The results discussed in this section suggest that schools cannot use basic statistical techniques and rely only on traditional demographic characteristics, previous student achievements, school attendance and behaviour in order to identify students at-risk. In other words, while it is true that graduation rates are lower, for instance, among African-American students or children in poor single-parent households, these variables are not enough to capture the variety of circumstances that lead students to halt their education. Similar poor results have been found in other studies on early warning indicators actually adopted in school districts (Deussen, Hanson and Bisht, 2017). Section Machine learning: results shows how high-dimensional data and ML algorithms can be combined in order to improve predictions.

Microeconomic foundation

So far, model performances have been evaluated by focusing on the recall rate. This section builds an economic model and introduce budget considerations to justify the use of the recall rate as a selection criterion. In this context, the optimization problem of the school (or school district officials) is the following: schools want to minimize the expected dropout rate subject to a budget constraint.[4] The goal is to correctly identify students at-risk in order to include them in a dropout prevention programme. This budget constraint takes into account the fact that the individual cost of the dropout prevention programme (τ) times the number of students enrolled in the programme has to be less or equal to total resources allocated to the programme (B).

The probability of dropping out $p(s_i, t_i)$ is defined as a function of the student's type (s_i) and the treatment (t_i), where the treatment is the dropout prevention programme. For simplicity, it is assumed that $s_i \epsilon \{0, 1\}$. In other words, there are two types of students: students at risk of dropping out ($s_{\mathrm{i}} = 1$) and students not at risk ($s_{\mathrm{i}} = 0$). The probability function $p(s_i, t_i)$ should satisfy certain properties:

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

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

[4] This objective function is consistent with goals set by federal and state legislations 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).

$$p\left(1,t\right)\geqslant 0. \tag{3.3}$$

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

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

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 students at-risk. Finally, condition (3.5) implies decreasing returns to scale, thus it is optimal to allocate resources equally among students at-risk

However, schools do not directly observe students who at risk, but rather only a signal, i.e. a predicted probability of dropping out provided by the algorithm. Given this signal, schools need to decide how many and which students to include in a dropout prevention programme in order to minimize the dropout rate. Therefore, using the notation introduced in section Technical considerations, the school optimization problem becomes:

$$\begin{aligned}&\min\{n_1[(1-\varphi)p(1,0)+\varphi p(1,t)]\}\\&\text{s.t.}\tau t[wr_1+c_1]\leqslant B\end{aligned}$$

where the objective function is the weighted sum of the number of students who end up dropping out and are not treated, plus those who are treated, each multiplied by the probability of dropping out given the treatment. As defined in section Technical considerations, φ is the recall rate, while $n_1$is the number of students who drop out. The cost of the programme in the budget constraint depends instead on the students which have been – both correctly and incorrectly – assigned to the treatment.

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

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

This linear function satisfies conditions (3.1)–(3.4). From this, it follows that the objective function becomes (excluding the constant n_1):

$$\begin{aligned}&\min\{(1-\varphi)*1+\varphi*0\}\\&\text{s.t.}\tau[wr_1+c_1]\leqslant B\end{aligned}$$

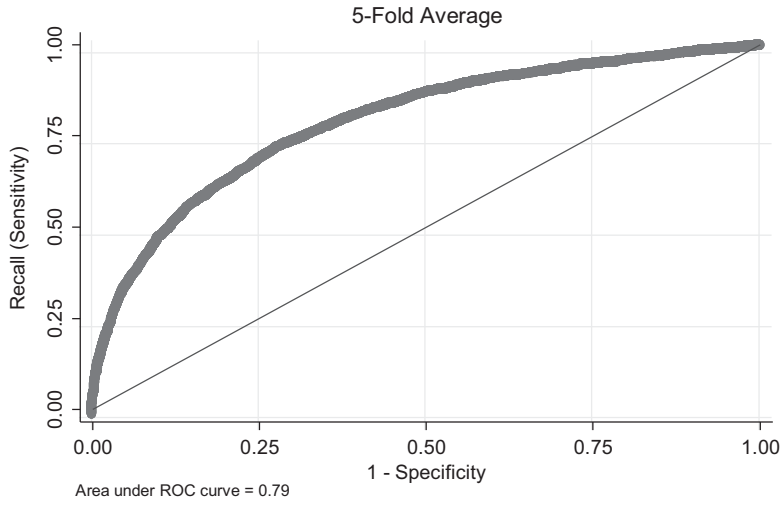

Figure 1. Logit ROC curve*Note:* Area under ROC curve = 0.79.

This is equivalent to maximizing the recall rate subject to a budget constraint. Therefore, this simple model provides an economic justification for using the recall rate as the criterion when tuning the ML algorithms through cross-validation and when comparing performances among them. It is worth emphasizing again that the additional advantage of using the recall rate rather than accuracy in this context is that it counteracts the negative effects of class imbalance, i.e. of having a relative small proportion of students not graduating from high school (Kuhn and Johnson, 2013).

Using the recall rate as the criterion can also be justified by imposing different functional forms on $p(s_i, t_i)$. For instance, the curvature imposed by the positive second derivative of $p(s_i, t_i)$ (Condition 3.5) can be taken into account by assuming the following functional form:

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

This would lead to an equivalent optimization problem:

$$\begin{gathered}\min\left\{1-\frac{\varphi}{2}\right\}\\ s.t.\tau[wr_1+c_1]\leqslant B\end{gathered}.$$

More generally, as long as $p(1,0)>p(1,t)$, the school objective function is equivalent to maximizing the recall rate.

A straightforward implementation of the above procedure can be applied to the Logit model discussed in section A Basic model (Table 1 Model 1). Previously, the threshold to estimate dropout status given predicted probabilities has been set at 0.5. However, it is possible to 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.[5] Ideally, a school would like to be as high as possible on the

[5] The Online Appendix A.2 provides a detailed explanation of how Figure 1 and Table 2 in this section have been computed.

TABLE 2

Optimal threshold (5-fold average)

Cost per student		*Overall Budget*		
		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%

Note: This table shows how the optimal accuracy and recall rate change given different combination of total budget and cost per student of a hypothetical effective high school dropout intervention programme.

y-axis, but the selected point cannot be too much on the right of the *x*-axis otherwise the programme exceeds the resources available. Indeed, after estimating the individual probability of dropping 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 high school graduates. Instead of using the area under the ROC curve as main criterion to compare algorithms as in Bowers *et al.* (2013), this section provides a theoretical model to justify the selection of the optimal point on the ROC curve.

Quite interestingly, the use of alternative cutoffs for the predicted probabilities is one of the strategies suggested to tackle class imbalance (Kuhn and Johnson, 2013). Therefore, this procedure not only adapts algorithms to the school objective function, but it also addresses the issues due to the low ratio between high school dropouts and graduates.

Table 2 shows how the optimal accuracy and recall rates change as schools vary the cost per student and the overall budget of the programme.[6] As a result, policy-makers can follow this procedure to choose the most efficient algorithm and tune its parameters in order to treat as many students at-risk as possible subject to their budget constraints. It is 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

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

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 programme does eventually exceed the resources initially allocated to it.

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.

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 a 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 x on y. Supervised ML are applied in section Machine learning: results to predict high school dropouts, while unsupervised ML are used in section IV 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, after including higher order terms and some interaction terms between the most important predictors, the number of independent 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 a Logit model. Adding too many variables to these models would lead to overfitting. 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]

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

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

[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, behaviour, sense of school belonging, 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).

Since LASSO 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). The key assumption is that the data generating process is sparse, where only a small subset of variables is assumed to have high predictive power. This may not be realistic in some economic applications (Giannone, Lenza and Primiceri, 2017).

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

$$\hat{\beta}(C) = \underset{\beta \epsilon \mathbb{R}^k}{\operatorname{argmin}} C_1 \left[\sum_{i=1}^{n} y_i \max\{0, 1 - K_i'\beta\} + (1 - y_i)\max\{0, K_i'\beta - 1\} \right] + \beta_2.$$

Here C_1 is the penalization parameter. Although kernel functions allow SVM to be extremely flexible, this comes at the cost of interpretability. The most common kernel is the Gaussian one, although the sigmoid kernel has also been considered in the empirical application.

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

$$K_{\text{sigmoid}}\left(x_1, x_2\right) \equiv \tanh\left(\theta + \gamma x_1' x_2\right)$$

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, misclassified observations in the previous classifier are given larger weights. Indeed, a simple version of Boosting can be illustrated by the AdaBoost algorithm:

1. Initially assign the same weight $1/n$ to all observations
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 can lead to higher performances than a single, more complex, algorithm such as Logit.

An example of a simple classifier often used within Boosting is a regression tree. This algorithm 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 thought of as a series of if-then statements, and it can be represented by a graph that looks like a tree. For instance, the observation may be divided into two groups: whether students have GPA below 2.0 or not. Then, those who satisfy this condition could be further split according to whether they are taking math in 9th grade, and so on. 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, Hastie and Tibshirani (2000) since these authors were able to reinterpret it in a likelihood framework, thus making it comparable to the objective function of an OLS or a Logit model. Boosting have been found to have superior performances than other ML algorithms in many simulations (Bauer *et al.*, 1999, 1999) and has already been used by Chalfin *et al.* (2016) in their work on predicting police hiring. Furthermore, ensemble-based methods such as Boosting have been shown to be effective in the presence of class imbalances (Chawla, 2010).

The next section presents results from these different algorithms because they offer different combination of interpretability and flexibility. Post-LASSO is easily interpretable since it just selects a subset of variables to use as predictors in an OLS model. The contribution of each variable is easily understood. On the other hand, SVM and Boosting are among the most flexible algorithms because they are able to fit an extremely large variety of functional forms. At the same time, they are 'black boxes' which do not provide detailed information on how the inputs have been combined, and thus lack transparency.

As discussed 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 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 *ex ante* how the variables interact among themselves: the algorithm selects the optimal combination with the highest predictive power.

Some other studies have used principal component analysis as a preliminary step to combine several variables into a few indicators to use them as predictors in a Logit model (Adelman *et al.*, 2018). 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 with the most predictive power, especially 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.

Machine learning: results

Table 3 reports the 5-fold out-of-sample performances of all the ML algorithms introduced in the previous section. 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

TABLE 3

ML (5-fold average)

	Algorithm	*Inputs*				*Performances*		
		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: This table reports out-of-sample performances of different models estimating the probability that a student drops out of high school. *Individual* indicates that the algorithm uses as inputs all the relevant variables from the student and parent questionnaires. *School* refers to inputs from the teachers, counsellor 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 are included in the final Logit model.

sample subject to a minimum accuracy rate (0.89, thus similar to the accuracy of the basic models in Table 1). As discussed in section Microeconomic foundation, 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.

As already mentioned in section Machine learning: brief introduction, 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 a Logit (Model 4) specification. As reported in Table 3, Post-LASSO algorithms manage to increase the recall rate up to 23%. Compared to the basic model, this is almost an eight percentage points increase, or a 51% improvement, while maintaining a comparable accuracy rate. The magnitude of these gains is substantial when interpreted at a national scale. The students interviewed in the HSLS:09 are representative of more than 4.1 million 9th grader in the U.S. Of these, around 483,270 ended up dropping out of high school. Therefore, each percentage point improvement in the recall rate implies that around 4,830 additional students would be correctly identified as at risk of dropping out. 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.

These out-of-sample performances are rather precise. For the Logit Post-LASSO (Model 4 Table 3), recall rates in the five folds used during cross-validation range between 19.3% and 27.9%. Therefore, the recall rate for this ML algorithm is higher than the ones obtained in Table 1 using a limited set of predictors not only on average, but even when every single fold is considered. Following Kuhn and Johnson (2013), it is also possible to repeat the 5-fold CV procedure multiple times and then use the different estimates of the recall rate in order to compute confidence interval and measure the prediction uncer-

[9] ML was not actually expected to provide perfect predictions. Indeed, as already mentioned, in order to allow schools enough time to identify students at-risk and target them with appropriate interventions, all predictors were collected in 9th grade. The implicit drawback 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.

tainty more formally. As shown in Table A1, repeating the 5-fold CV five times produces a confidence interval of [0.168; 0.286], while repeating it 10 times restricts the interval up to [0.173; 0.277], thus increasing the precision of the estimates while maintaining performances always superior to those in Table 1.

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).

The above performances of the ML algorithms are in line with a few previous case studies and extend the work done by other researchers using traditional econometric techniques to predict high school dropout in both developed and developing countries (see for instance Rumberger and Lim, 2008; Bowers *et al.*, 2013; Adelman *et al.*, 2018). There have been some very 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 3 months later. Aguiar *et al.* (2015) introduced ML to predict which students are at risk of dropping out in a US 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 in the introduction, 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 US nationally representative data set, thus reducing the external validity concerns raised for local analysis.

Robustness checks and extensions

Different objective function

The algorithms presented in Table 3 are extremely flexible and can be adapted to different objective functions. For instance, if Logit Post-LASSO (Model 4) is calibrated in order to maximize the area under the ROC curve, it reaches an AUC of 0.81, while maintaining an accuracy of 89.8%, as well as a recall rate of 18.2%. Similarly, if the same algorithm is calibrated to maximize the accuracy rate, it obtains a similar rate of the one in Table 1 (89.9%), but at the same time the AUC and recall rate are higher than the ones obtained with the basic model (18.3% and 0.81 compared to 15.2% and 0.80).

These variations demonstrate how these high-dimensional techniques can dominate basic models under many performance criteria. 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 justified goodness-of-fit measure as discussed in section Microeconomic foundation.

Additional specifications and algorithms

Table A2 in the Online Appendix reports results from additional algorithms and specifications. First, including school fixed effects to SVM (Model 1) or Boosting (Model 2) does not lead to better performances than those obtained from Post-LASSO Logit (Table 3). Similarly, including additional interaction terms or school fixed effects to Post-LASSO OLS does not provide improvements in performances (Models 3 and 4 Table A2).

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 Machine learning: brief introduction, it has been decided to only report results for these three algorithms since there are among the most popular ones and they have been shown to have superior performances in many simulations. Moreover, their calibration is not extremely time-consuming, thus avoiding the risk that such techniques may be computationally infeasible for schools given their limited technological equipment. Indeed, more advanced algorithms may still be hard to scale up, even for big companies (Johnston, 2012), or extremely difficult to code, which is the reason behind the very high prizes – often reaching $1 million (Netflix, 2009) - offered in machine learning competitions.

For the sake of completeness, Table A2 reports the out-of-sample performances for a Ridge regression (Model 5), as well as a more general Elastic Net (Model 6). These algorithms are describe in Friedman, Hastie and Tibshirani (2010). As expected given their objective functions similar to LASSO, both algorithms do not perform better than Post-LASSO Logit.

Additional or alternative inputs

The main analysis does not include whether the 9th grader participated in certain programmes that may have affected his or her probability of finishing high school. While it is true that between 2009 and 2013 some of these students received some treatment to reduce their risk of dropping out, this holds across all specification, even the ones without ML in Table 1. Thus, the existence of these programmes does not undermine the conclusion that ML algorithms provide superior performances. Furthermore, it is unclear whether these variables should be included as inputs in the algorithm. Adding them may increase the predictive power of the algorithms, but these gains would be obtained by predicting students at risk of dropping out by using participation in high school dropout prevention programmes, which may seem recursive. Model 7 in Table A2 replicates Model 4 in Table 3, but it also includes whether the 9th graders participated in the following programmes: Talent Search, Upward Bound, Gear Up, Advancement Via Individual Determination (AVID), and Mathematics, Engineering, Science Achievement (MESA). Section A.3.2 in the Online Appendix describes these programmes. As expected given the evidence on their (limited) effectiveness, including these variables as inputs does not substantially affect the performances of the algorithm.

For a few variables, information has been obtained from the first or second follow-up interviews because of the lower number of missing values than the baseline survey. Additional questions were asked to the students in the follow-up interviews if their parents

had not responded in the baseline survey (or vice versa). Model 8 in Table A2 replicates Model 4 in Table 3 while using only information from the baseline survey regarding student's ethnicity and language, household income, household size, as well as mother's and father's educational level, employment and occupation. Most of these variables tend to be time-invariant, so it is not surprising that the out-of-sample performances of the Post-LASSO Logit do not change substantially. This result is also reassuring since it supports the idea that schools can already identify students at-risk in 9th grade with sufficient precision.

The HSLS:09 contains some variables that have high predictive power, but are usually unavailable to schools or might be difficult to obtain. Therefore, in order to estimate the algorithms under more realistic data scenario, Model 9 in Table A2 estimates the same Post-LASSO Logit model but with a restricted set of inputs. In particular, the list of regressors no longer include information about students' expectations on their future education and career; their relationships with parents and peers; their time management, behaviour and self-perception; their parents' expectations, level of support, involvement and behaviour. Despite this limited set of independent variables, the out-of-sample performances of the algorithm do not change substantially. As discussed in section Pivotal variables, LASSO mostly selects variables available from academic transcripts or other administrative data. Even if these additional behavioural and psychological variables were powerful predictors, it seems that they can be substituted with information contained in other available data, thus not impacting the performances of the algorithm. It is also worth noting that schools often have more detailed information regarding their teacher body than the HSLS:09, thus they might actually reach even higher performances by including these teacher characteristics in their algorithms.

Coding outcome variable

As discussed in section Outcome variable, 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 (re-takers are not counted as dropouts). 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 actually improves the recall rate (even if it reduces the sample size to around 16,400 observations). As shown in Model 10 in Table A2, estimating the same Logit Post-LASSO algorithm as Model 4 in Table 3 for this alternative outcome variable leads to a recall rate of 35.7%. For comparison, the recall rate of a Logit model as the one in Table 1 for the same alternative outcome variable reaches a recall rate of 28.2%.

Heterogeneity across regions

As already discussed, one of the differences between this analysis and previous studies is the use of a recent U.S. nationally representative data set. As a result, it is possible to argue that ML techniques would lead to substantial improvements in identifying students at risk of dropping out across the entire nation, not only in certain localities or context. However, there is the risk that the algorithm may correctly identify students at-risk only in certain regions. In line with this concern, there is some variability in the recall rate across regions.

The recall rate ranges from 19.2% in the Northeast, to 28.3% in the Midwest.[10] However, the gains from using ML algorithms are not concentrated only in one region, and even the lowest recall rate is higher than the ones from basic models (Table 1).

Some practitioners may also be interested in examining the performances of these algorithms when estimated within certain regions. Indeed, these techniques may be initially implemented only in certain school districts. While there are no methodological differences and the same algorithms can be easily re-estimated using only data from certain U.S. sub-regions or states, the sample size is considerably reduced. This may not be an issue when using large administrative data, but it limits the ability of the algorithms to disentangle noise from signal in this specific exercise given the finite dimension of the HSLS:09.

Model 11 in Table A2 replicates the Logit Post-LASSO model reported in Table 3 (Model 4) using observations from U.S. states in the South. This region has been chosen since it has some of the states with the lowest high school graduation rates – such as Georgia and Louisiana – and because it has a relatively large sample size. The recall rate is still higher than a simple Logit model as the one reported in Table 1 estimated on the same sub-sample (17.4% vs. 13.3% respectively). Nevertheless, such recall rate remains rather low and relatively far from the ones reported obtained from the full sample (Table 3). Therefore, this result confirms that the gains from using ML are considerably larger when these techniques are applied to very large datasets.

Including equity in school objective functions

Recently, there have been some concerns about the hidden biases within ML algorithms and the ethical consequences of their diffusion (Sweeney, 2013). However, this issue is limited in this context since the goal of this paper is only to provide schools with better information about their students. The algorithms are not aimed at selecting which courses should each student take. More generally, it is worth emphasizing that algorithms can have biases, but these can often be easily detected and eliminated, while the same cannot be said about the widespread biases in human evaluations and decisions.

With these caveats in mind, it might still be interesting to discuss whether it might be socially desirable to exclude certain variables such as race or gender – or another set of variables collinear with them – from the list of inputs in order to avoid biases in the algorithms. For instance, one might be worried that a ML algorithm might identify too many (or too few) black students as at risk of dropping out because of stereotypes and past discriminations reflected in the training sample. Alternatively, due to the higher dropout rate among Hispanics and African-American students, schools may prefer to target these groups.

These equity concerns can be easily included in the main theoretical framework introduced in section Microeconomic foundation. Define $\omega(S)$ as:

$$\omega(S) = -n_1[(1-\varphi)p(1,0) + \varphi p(1,t)],$$

where S is the set of students identified as at risk of dropping out and admitted to the dropout prevention programme, i.e. $S = wr_1 + c_1$. The objective of the school is to maximize $\omega(S)$

[10] Recall rates computed from the predicted probabilities across the 5 folds of the Logit Post-LASSO algorithm reported in Model 4 Table 3.

(subject to the budget constraint). It is possible to include preferences regarding certain observable characteristics – e.g. gender or race – in the school optimization problem:

$$\max \omega(S) + \vartheta(S)$$

where $\vartheta(S)$ is monotonically increasing in the number of students in S who belongs to the preferred category. In other words, this new school objective function includes an efficient component (minimize the number of dropouts) and an equitable component (prioritize certain categories). In this case, schools should still use all available information, including gender and race, in order to obtain accurate predictions for each students (Kleinberg *et al.*, 2018). The above equity considerations can then be satisfied by selecting a different cutoff for each group to admit students in the dropout prevention programme.

If, for instance, schools are interested in focusing on male black students, the algorithms discussed in the previous sections can be easily adapted in this context by using a lower threshold to convert predicted probabilities into predicted outcomes (graduate/dropout) for these students, thus affecting the racial composition of the set S. In other words, instead of using 0.5 for all observations as typically done in most algorithms, one can select one threshold for male black students and a higher one for all the other students in order to achieve the desired racial composition, as well as to respect the budget constraint by not including too many students in the programmes. The key takeaway is that, even when schools care about equity, it is optimal to incorporate any observable variable as input in the algorithm.

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 measure the influence of a variable in the boosted regression model estimated in Table 3 (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.

One can look 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 13 of them have been selected more than three times. Table 4 lists these 13 predictors along with the number of boosted regressions they have been used in, and the 5-fold average influence.[11]

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. As highlighted in the previous literature, past academic performances, attendance and school behaviour are indeed important predictors. In particular, GPA in 9th grade is always selected and its average influence is rather high.

[11] The ranking is similar if variables are sorted based on the average influence. Table A3 in the Online Appendix lists the 33 predictors which have been selected at least 2 times.

TABLE 4

Variables selected by boosting

Predictors	*Count*	*Influence*
GPA in 9th grade	5	39.7
Born in 1993 (most students were born in 1994–95)	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

Note: This table lists the variables selected by Boosting (Table 3 Model 2) at least three times in the 5-fold estimation. The influence measures the average overall contribution of each variable to the log-likelihood function. Such values are standardized between 0 and 100.

Despite these commonalities, the list 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 also predicts dropout. Contrary to the wide-spread belief that the ABC (Attendance, Behaviour, Course grades) system is able to capture the impact of family characteristics (Rumberger *et al.*, 2017), number of household members is often selected, highlighting the additional predictive power of household background information. Finally, subjective expectations 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 emphasized the importance of combining such variables with additional – carefully selected – predictors and to use advanced techniques to optimally combine them. 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 3, using few indicators cannot match the performances obtained with a larger set of variables

A similar exercise can be conducted with LASSO. In particular, one can look at the top predictors (around 20–26 in each fold) selected by LASSO to generate the two-way interaction terms in Table 3 Model 6. Among these selected inputs, Table 5 reports the list of variables picked in at least three of the five folds. Several variables appear in both Tables 4 (Boosting) and 5 (LASSO): GPA, year of birth, math test score, no math or science course taken in 9th grade, school transfers, attendance, behaviour, and expectations about school attainments. It is remarkable that both algorithms select these variables. This supports the conclusion of their high predictive power. In addition, LASSO frequently selected a few

TABLE 5

Variables selected by LASSO

Predictors	*Count*
Born in 1993 (most students were born in 1994–95)	5
Born in 1995	5
HSLS:09 Math test score	5
No math courses taken in 9th grade	5
No science courses taken in 9th grade	5
GPA in 9th grade	5
9th grader very sure that he/she will graduate from high school	5
Public School	5
Private School	5
Whether 9th grader has ever been suspended or expelled	5
Does not plan to enroll in college after high school	4
Principal reporting student drop out not a problem	4
9th graders has never changed schools since kindergarten	4
Parent reporting no difficulty by 9th grader with behaviour 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 9th grader will at most attain high school	4
GPA for all academic 9th grade courses	3
9th grader thinks he/she will at most attain high school	3
9th grader did not repeat 2nd grade	3
9th grader spend <1 hour/day on extracurricular activities	3
9th grader was in 9th grade in the previous academic year	3

Note: This table lists the variables selected by LASSO to generate the two-way interaction terms includes in Model 6 Table 3. Only variables selected in at least 3 of the 5 folds have been included.

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 important to emphasize again that these variables are identified by ML algorithms as important predictors. This does not imply that changing these variables would lead to a reduction of school dropout rates. The aim of this analysis is to provide precise predictions, not causal inference. This does not reduce the contribution of the paper: both causality and prediction are relevant in this context since policy-makers are interested in identifying students at-risk, as well as understanding which variables can be affected to reduce their risk of dropping out.

To reiterate the argument discussed 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 final set of selected variables 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 the algorithm provides accurate predictions, which variables are chosen is irrelevant in this context given the absence of any causal interpretation. For instance, gender and ethnicity are – quite surprisingly - not used as main predictors by LASSO and Boosting, but

this does *not* imply that these factors are irrelevant in this context or that they would not be selected using a different training set. In other words, it is possible that the variables listed in Table 4 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.

It is also important to note that most of the predictors in Tables 4 and 5 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 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 only using 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 programme, 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.[12]

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 the Logit Post-LASSO model (Table 3 Model 4) is calibrated by excluding such a variable from the list of potential predictors, the algorithm still reaches 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 data set. 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 primarily designed to detect students at risk of dropping out, the above results suggest that schools can efficiently predict which students are going to drop out without having to rely on additional expensive tests, but by analysing available individual, family and school characteristics.

IV. Clustering predicted dropouts

Identifying students at-risk is only the first step. Next, schools have to design the appropriate programmes for them. However, as also emphasized 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 counselling may be more effective for students with discipline issues or problems at home.[13] In other words, this section acknowledges that high school dropout is

[12] 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.

[13] Income inequality may also play a role for individuals from low socio-economic background (Kearney and Levine, 2016): greater income gaps between those at the bottom and those at the middle of the income distribution may lead low-income students (especially boys) to drop out of high school due to a 'despair' effect – seeing the middle class as unattainable - rather than an aspirational effect. The number of students per school in the HSLS:09

TABLE 6

Clustering

Predictors	*No dropout*	*Group 1*	*Group 2*	*Group 3*	*Group 4*
Born in 1993 (most students were born in 1994–95)	0.03	0.22	0.63	0.13	0.65
HSLS:09 Math test score	0.47	0.31	0.31	0.32	0.33
No math courses taken in 9th grade	0.04	0.26	0.19	0.40	0.10
No science courses taken in 9th grade	0.06	0.31	0.25	0.59	0.14
GPA in 9th grade	0.63	0.25	0.27	0.21	0.31
9th 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
Whether 9th 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 high school	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
Parent reporting no difficulty by 9th grader with behaviour 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 participated in school fundraiser	0.39	0.11	0.10	0.14	0.39
Parent thinks 9th grader will at most attain high school	0.05	0.15	0.59	0.34	0.04
9th grader was in 9th grade in the previous academic year	0.04	0.23	0.33	0.43	0.39
Observations	20,340	630	110	120	100

Note: This table reports the summary statistics (mean) for each group identified by the hierarchical clustering algorithm: students identified as at risk of dropping out have been divided into four groups. The table also reports summary statistics for the group of students not predicted to drop out of high school. All variables have been rescaled between 0 and 1.

a multidimensional issue: different factors may lead students to halt their education. This is similar to the multidimensional approach advocated in poverty studies (Alkire and Foster, 2011). This section shows how students predicted to dropout can be divided into different subgroups using unsupervised machine learning.

The starting point is the prediction obtained using the Logit Post-LASSO algorithm in Table 3 (Model 4). In line with the results in section Pivotal variables, the same predictors selected by this algorithm at least in 3 of the 5 folds (Table 5) have been used to divide 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 6 shows the summary statistics for these

is too small to compute reliable statistics of within-school inequality, but school districts could incorporate such a measure – even a within-grade or within-class inequality index - in their algorithms.

predicted dropouts.[14] For comparison, the second column 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.[15] Group 1 is mainly composed by individuals with low attendance, behavioural issues, and lack of parental involvement. On the other hand, students in Group 2 were older than ta 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 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. This result emphasizes the importance of not pooling together all students at risk of dropping out. Placing well-behaved but academically weak students in a classroom side by side with students with suspension and low attendance records may actually results in negative externalities.

V. 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. It illustrates how Big Data and ML can be fruitfully applied in education to improved school performances by efficiently using all 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 students at-risk and draw from the existing literature to identify effective programmes to help them. Another advantage of these early predictions is that counsellors and teachers may suggest vocational careers to these vulnerable students (Goux, Gurgand and Maurin, 2017). Last but not least, following the growing literature on the pivotal role of information constraints in education (Hoxby and Turner, 2015), parents could be informed on whether their students are considered as at risk of dropping out.

[14] For simplicity, only the key variables have been reported Table 6. Summary statistics for the whole set of predictors are reported in the Online Appendix (Table A5)

[15] 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. Indeed, each group contains both correctly and incorrectly predicted dropouts: 48% of students in Group 1 did end up dropping out. The same percentage is 61% for Group 2, 66% for Group 3, and 43% for Group 4. section A.4 in the Online Appendix argues that students misclassified as predicted dropout are actually week students even if they graduated from high school, thus they would have still benefited from additional support.

Although using few indicators may be attractive, this paper highlights that this approach leads to extremely unreliable predictions. Schools have additional information, and their data sets are increasing exponentially over time thanks to new technologies. ML can help practitioners to efficiently use them. 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. Future research could also investigate whether alternative ML algorithms produce even larger gains in terms of prediction accuracy. 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, this study has showed not only that supervised ML can improve school predictions, but also that unsupervised ML can identify sub-populations among students at-risk. Therefore, schools may design the appropriate programme 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 that 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 benchmark 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. For instance, one can assume that there are two types of students - those who are at risk of dropping out, and those who are not - and that there is an effective treatment which can be provided by schools and which has a homogeneous impact on students at-risk. In other words, it is assumed that there is a dropout prevention programme which is able to equally reduce the probability of dropping out for all treated struggling students.[16] High-dosage tutoring is an example of a policy that can help these students (Fryer, 2017). The necessary pre-condition to implement this programme is to identify the students who need the treatment, i.e. those at risk of not graduating from high school. 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 programme. Schools need to know if a student belongs to the 'not at risk' category or to the 'at risk' one. 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 at a higher risk of dropping out among the student population, while unsupervised ML can divide these students into subgroups, and then scarce and expensive human resources can be invested to design the best intervention for these restricted set of students. Therefore,

[16] Note that this assumption does not require homogeneous treatment for the whole population, but only for the vulnerable students. In fact, the treatment may be completely ineffective for students who have high probability of graduating from high school.

even if current ML techniques are designed to provide accurate predictions, but they are often inappropriate to optimally allocate resources (Athey, 2017), they can still provide complementary tools for causal inference. Put differently, ML does not substitute traditional economic models and econometric estimations, but provides an additional technique to reinforce and strengthen those analyses.

Final Manuscript Received: October 2018

References

Adelman, M., Haimovich, F., Ham, A. and Vazquez, E. (2018). 'Predicting school dropout with administrative data: new evidence from guatemala and honduras', *Education Economics*, Vol. 26, pp. 356–372.

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B. and Addison, K. L. (2015). 'Who, when, and why: a machine learning approach to prioritizing students at risk of not graduating high school on time', *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*, Vol. 15, pp. 93–102.

Alkire, S. and Foster, J. (2011). 'Counting and multidimensional poverty measurement', *Journal of Public Economics*, Vol. 95, pp. 476–487.

Allensworth, E., Nomi, T., Montgomery, N. and Lee, V. E. (2009). 'College preparatory curriculum for all: academic consequences of requiring Algebra and English I for ninth graders in Chicago', *Educational Evaluation and Policy Analysis*, Vol. 31, pp. 367–391.

Athey, S. (2017). 'Beyond prediction: using big data for policy problems', *Science*, Vol. 355, pp. 483–485.

Aulck, L., Velagapudi, N., Blumenstock, J. and West, J. (2016). Predicting student dropout in higher education. *ArXivWorking Paper* 1606.06364.

Bauer, E., Kohavi, R., Chan, P., Stolfo, S. and Wolpert, D. (1999). 'An empirical comparison of voting classification algorithms: bagging, boosting, and variants', *Machine Learning*, Vol. 36, pp. 105–139.

Bedard, K. and Do, C. (2005). 'Are middle schools more effective? The impact of school structure on student outcomes', *Journal of Human Resources*, Vol. 40, pp. 660–682.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). 'High-dimensional methods and inference on structural and treatment effects', *Journal of Economic Perspective*, Vol. 28, pp. 29–50.

Bowers, A. J. and Sprott, R. (2012). 'Why tenth graders fail to finish high school: a dropout typology latent class analysis', *Journal of Education for Students Placed at Risk*, Vol. 17, pp. 129–148.

Bowers, A. J., Sprott, R. and 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', *The High School Journal*, Vol. 96, pp. 77–100.

Carnevale, A. P., Smith, N. and Strohl, J. (2013). *Recovery - Job Growth and Education Requirements through 2020*, Center on Education and the Workforce, Georgetown University, Washington, DC.

Carniero, P. and Heckman, J. J. (2003). 'Human capital policy', in Heckman J. J. and Krueger A. B. (eds), *Inequality in America: What Role for Human Capital Policies?*, Cambridge, MA: MIT Press, pp. 77–240.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig J. and Mullainathan S. (2016). 'Productivity and selection of human capital with machine learning', *American Economic Review*, Vol. 106, pp. 124–127.

Chawla, N. V. (2010). 'Data mining for imbalanced datasets: an overview', in Maimon O. and Rokach L. (eds), *Data Mining and Knowledge Discovery Handbook*, Boston: Springer, pp. 853–867.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2015). 'The aftermath of accelerating algebra: evidence from district policy initiatives', *Journal of Human Resources*, Vol. 50, pp. 159–188.

Cook, P. J., Dodge, K., Farkas, G., Fryer, Jr R. G., Guryan, J., Ludwig, J., Mayer, S., Pollack, H. and Steinberg, L. (2014). 'The (surprising). 'efficacy of academic and behavioral intervention with disadvantaged youth: results from a randomized experiment in Chicago', *National Bureau of Economic Research*, Vol. 19862, pp. 1–59.

Cortes, K. E., Goodman, J. S. and Nomi, T. (2015). 'Intensive math instruction and educational attainment: long-run impacts of double-dose algebra', *Journal of Human Resources*, Vol. 50, pp. 108–158.

Deussen, T., Hanson, H. and 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 (REL 2017–261)*, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northwest, Washington, DC. Available at: Retrieved from http://ies.ed.gov/ncee/edlabs.

District Performance Office. (2017). *2015–16 School Progress Report*, The School District of Philadelphia, Philadelphia.

Ekowo, M. and Palmer, I. (2016). *The Promise and Peril of Predictive Analytics in Higher Education*, New America, Washington, DC.

Friedman, J. (2001). 'Greedy function approximation? A gradient boosting machine', *Annals of Statistics*, Vol. 29, pp. 1189–1232.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). 'Additive logistic regression: A statistical view of boosting', *Annals of Statistics*, Vol. 28, pp. 337–407.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software*, Vol. 33, pp. 1–22.

Fryer, R. G. (2017). 'The production of human capital in developed countries: evidence from 196 randomized field experiments', in Duflo E. and Banerjee A. (eds), *Handbook of Field Experiments*, Amsterdam: North Holland, pp. 95–322.

Giannone, D., Lenza, M. and Primiceri, G. E. (2017). *Economic Predictions with Big Data: The Illusion of Sparsity*, Working Paper.

Goux, D., Gurgand, M. and Maurin, E. (2017). 'Adjusting your dreams? High school plans and dropout behaviour', *The Economic Journal*, Vol. 127, pp. 1025–1046.

Guenther, N. and Schonlau, M. (2016). 'Support vector machines', *Stata Journal*, Vol. 16, pp. 917–937.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. Springer, New York.

Heckman, J. J., Humphries, J. E. and Mader, N. S. (2011). 'The GED', in Hanushek E. A. , Machin S. and Woessmann L. (eds), *Handbook of the Economics of Education*, North-Holland: Elsevier, pp. 423–484.

Heckman, J. J. and LaFontaine, P. A. (2010). 'The American high school graduation rate: trends and levels', *Review of Economics and Statistics*, Vol. 92, pp. 244–262.

Heckman, J. J. and Rubinstein, Y. (2001). 'The importance of noncognitive skills: lessons from the GED testing program', *American Economic Review*, Vol. 91, pp. 145–149.

Hoxby, C. and Turner, S. (2015). 'What high-achieving low-income students know about college', *American Economic Review*, Vol. 105, pp. 514–517.

IES. (2016). *Public High School Graduation Rate Reaches New High, But Gaps Persist*, U.S. Department of Education, Institute of Education Sciences, Washington, DC.

Ingels, S. J., Pratt, D. J., Herget, D., Bryan, M., Fritch, L. B., Ottem, R., Rogers, J. E. and Wilson, D. (2015). *High School Longitudinal Study of 2009 (HSLS:09). 2013 Update and High School Transcript Data File Documentation*, U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC.

Ingels, S. J., Pratt, D. J., Herget, D. R., Burns, L. J., Dever, J. A., Ottem, R., Rogers, J. E., Jin, Y. and Leinwand, S. (2011). *High School Longitudinal Study of 2009 (HSLS:09), Base-Year Data File Documentation*, U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC.

Ingels, S. J., Pratt, D. J., Herget, D. R., Dever, J. A., Fritch, L. B., Ottem, R., Rogers, J. E., Kitmitto, S. and Leinwand, S. (2014). *High School Longitudinal Study of 2009 (HSLS:09), Base Year to First Follow-Up Data File Documentation*, U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, Washington, DC.

Johnston, C. (2012). Netflix Never Used Its $1 Million Algorithm Due To Engineering Costs, *Wired* (April) Available at: https://www.wired.com/2012/04/netflix-prize-costs/.

Kearney, M. S. and Levine, P. B. (2016). 'Income inequality, social mobility, and the decision to drop out of high school', *Brookings Papers on Economic Activity*, Vol. Spring, pp. 333–380.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015). 'Prediction policy problems', *American Economic Review*, Vol. 105, pp. 491–495.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Rambachan, A. (2018). 'Algorithmic fairness', *AEA*, Vol. 108, pp. 22–27.
Knowles, J. (2015). 'Of needles and haystacks: building an accurate statewide dropout early warning system in Wisconsin', *JEDM: Journal of Educational Data Mining*, Vol. 7, pp. 1–52.
Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*, Springer, New York.
Luca, M., Kleinberg, J. and Mullainathan, S. (2016). 'Algorithms need managers, too', *Harvard Business Review*, Vol. 104, pp. 96–101.
Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I. and 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 Research Notes*, Vol. 4, pp. 299.
McKenzie, D. and Sansone, D. (2017). 'Man vs. machine in predicting successful entrepreneurs: evidence from a business plan competition in Nigeria', *World Bank WP*, Vol. 8271, pp. 5.
Mullainathan, S. and Spiess, J. (2017). 'Machine learning: an applied econometric approach', *Journal of Economic Perspective*, Vol. 31, pp. 87–106.
Murnane, R. J. (2013). 'U.S. high school graduation rates: patterns and explanations', *Journal of Economic Literature*, Vol. 51, pp. 370–422.
Netflix. (2009).Netflix Prize, Retrieved (January 1, 2017). Available at: http://www.netflixprize.com/.
Ng, A. (2016). Machine Learning, *Coursera*. Retrieved (January 1, 2016). Available at: https://www.coursera.org/learn/machine-learning.
O'Cummings, M. and Therriault, S. B. (2015). *From Accountability to Prevention? Early Warning Systems Put Data to Work for Struggling Students*, American Institutes for Research, Washington, DC.
OECD. (2016). *Education at a Glance (2016): OECD Indicators*, OECD Publishing, Paris.
Office of Innovation & Improvement. (2016). *Investing in Innovation Fund (i3)*, U.S. Department of Education, Washington, DC.
Oreopoulos, P. (2007). 'Do dropouts drop out too soon? Wealth, health and happiness from compulsory schooling', *Journal of Public Economics*, Vol. 91, pp. 2213–2229.
Oreopoulos, P. and Salvanes, K. G. (2011). 'Priceless: the nonpecuniary benefits of schooling', *Journal of Economic Perspective*, Vol. 25, pp. 159–184.
Rodriguez-Planas, N. (2012). 'Longer-term impacts of mentoring, educational services, and learning incentives: evidence from a randomized trial in the United States', *American Economic Journal: Applied Economics*, Vol. 4, pp. 121–139.
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. and Tuttle, C. (2017). *Preventing Dropout in Secondary Schools (NCEE 2017-4028)*, U.S. Department of Education, National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, Washington, DC. Available at: https://whatworks.ed.gov.
Rumberger, R. W. and Lim, S. A. (2008). *Why Students Drop Out of School: A Review of 25 years of Research*, UC Santa Barbara, Santa Barbara.
Sara, N. B., Halland, R., Igel, C. and 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*., (November 2014), 22–24.
Schonlau, M. (2005). 'Boosted regression (boosting): an introductory tutorial and a Stata plugin)', *Stata Journal*, Vol. 5, pp. 330–354.
Schwerdt, G. and West, M. R. (2013). 'The impact of alternative grade configurations on student outcomes through middle and high school', *Journal of Public Economics*, Vol. 97, pp. 308–326.
Shields, K. A., Cook, K. D. and Greller, S. (2016). *How Kindergarten Entry Assessments are Used in Public Schools and How They Correlate with Spring Assessments (REL 2017–182)*, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands, Washington, DC. Available at: http://ies.ed.gov/ncee/edlabs.
Subrahmanian, V. S. and Kumar, S. (2017). 'Predicting human behavior: the next frontiers', *Science*, Vol. 355, pp. 489.
Sweeney, L. (2013). 'Discrimination in online ad delivery', *Communications of the ACM*, Vol. 56, pp. 44–54.

U.S. Department of Education. (2009). *Race to the Top Program: Executive Summary*, U.S. Department of Education, Washington, DC.
Varian, H. R. (2014). 'Big data? New tricks for econometrics', *Journal of Economic Perspective*, Vol. 28, pp. 3–28.
De Witte, K., Cabus, S., Thyssen, G., Groot, W. and Van Den Brink, H. M. (2013). 'A critical review of the literature on school dropout', *Educational Research Review*, Vol. 10, pp. 13–28.
Witten, D. M. and Tibshirani, R. (2010). 'A framework for feature selection in clustering', *Journal of American Statistical Association*, Vol. 105, pp. 713–726.
Zajacova, A. (2012). 'Health in working-aged Americans: adults with high school equivalency diploma are similar to dropouts, not high school graduates', *American Journal of Public Health*, Vol. 102, pp. 284–290.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix S1. Online Appendix.