CRISTIAN CRESPO
London School of Economics and Political Science, London, UK

# Two Become One: Improving the Targeting of Conditional Cash Transfers with a Predictive Model of School Dropout

**ABSTRACT** This paper offers a methodology to improve targeting design and assessment when two or more groups need to be considered, and trade-offs exist between using different targeting mechanisms. The paper builds from the multidimensional targeting challenge facing conditional cash transfers (CCTs). I analyze whether a common CCT targeting mechanism, namely, a proxy means test (PMT), can identify the poor and future school dropouts effectively. Despite both being key target groups for CCTs, students at risk of dropping out are rarely considered for CCT allocation or in targeting assessments. Using rich administrative data sets from Chile to simulate different targeting mechanisms, I compare the targeting effectiveness of a PMT and other mechanisms based on a predictive model of school dropout. I build this model using machine learning algorithms. Using two novel metrics, I show that combining the outputs of the predictive model with the PMT increases targeting effectiveness except when the social valuation of the poor and future school dropouts differs to a large extent. More generally, public officials who value their key target groups equally may improve policy targeting by modifying their allocation procedures.

*JEL Codes:* I32, I38, I39

*Keywords:* Multidimensional targeting, conditional cash transfers, school dropout prediction, machine learning

Conditional cash transfers (CCTs) have become a favored social policy in developing nations. The use of these programs has expanded rapidly, from a few countries in the late 1990s to more than sixty by 2014 (Honorati, Gentilini, and Yemtsov, 2015). Although the stated objectives of CCTs vary, these schemes generally seek to reduce the incidence and depth of poverty (Handa and Davis, 2006) and provide a minimum consumption floor to poor households (Fiszbein and Schady, 2009).

**ACKNOWLEDGMENTS** I thank the chief editor for insightful comments and suggestions. I am also grateful to Stephen Jenkins, Grace Lordan, Emla Fitzsimons, and Emilia del Bono for their detailed feedback.

Targeting is a crucial element in the design of CCTs. These programs have tended to allocate their benefits primarily or "rather narrowly" to the poor (Fiszbein and Schady, 2009, p. 7). The more resources that are directed toward this key target group, the more likely a CCT is to achieve its goal of poverty reduction. This explains why evaluations of their targeting focus primarily on whether the CCTs have been given to those who live in poverty (Maluccio, 2009; Robles, Rubio, and Stampini, 2015; Skoufias, Davis, and de la Vega, 2001; Stampini and Tornarolli, 2012).[1]

Targeting low-income households or individuals makes sense not only for CCTs but for a wide range of social programs. Correspondingly, assessing social policy–targeting mechanisms in terms of their ability to find this target group is a widespread practice. For example, Coady, Grosh, and Hoddinott (2004) evaluate the pro-poor targeting performance of 122 social programs from forty-eight countries. Similarly, Grosh and Baker (1995) assess whether proxy means tests (PMTs) provide useful information on income for targeting social programs in three countries in Latin America.[2] Social policy targeting has been implicitly associated with finding the poor and alleviating poverty.

However, policymakers often need to allocate a single policy to different target groups. They might thus need to target their social programs using more dimensions than only income or poverty. For example, since many CCTs are provided only if children or adolescents are enrolled in school, an additional purpose of most CCTs is to increase school enrollment (Handa and Davis, 2006). To maximize the likelihood of achieving this goal, CCTs should also be delivered to a differently defined target group, namely, students with the highest risk of dropping out of primary or secondary school.[3]

1. Stampini and Tornarolli (2012) provide targeting assessments for thirteen countries in Latin America. They show that the expansion of CCTs on the continent led to increased inclusion of the poor, such that by 2010 the three largest programs (in Colombia, Mexico, and Brazil) had achieved poor coverage rates near 50 percent. However, this was accompanied by growing levels of nonpoor leakage (that is, the proportion of CCT recipients who are not poor). On average, leakage increased by 0.46 percentage points for each additional point in poor coverage.

2. Proxy means tests are one of the most common targeting mechanisms used for CCTs in Latin America (Fiszbein and Schady, 2009; Stampini and Tornarolli, 2012). In these systems, information correlated with income is used in a statistical formula to proxy income, using data that are easily observable by public officials (Coady, Grosh, and Hoddinott, 2004; Grosh and Baker, 1995).

3. CCTs are rarely assessed in terms of their ability to reach those who are more likely to drop out of school. My paper addresses this gap in the CCT literature. Analyzing the targeting effectiveness in reaching students at risk of dropping out of school is different from assessing the impact on school dropout. The former assesses whether the target group is (or would be)

In Latin America, the number of CCT beneficiaries overtook the poor population in 2006. The massive rise in CCT recipients on the continent has raised a debate in the literature about whether these schemes have gone too far (Stampini and Tornarolli, 2012). Although the relationship between CCT recipients and the population in poverty varies by country, the outreach of many CCT programs has exceeded the population living in poverty. In this context, identifying new poor beneficiaries has become harder. Successfully targeting multiple groups in CCTs can thus be especially relevant in these countries.

Assessing the capacity of CCTs to reach potential school dropouts is very important. If the targeting mechanism used by a CCT is not an accurate predictor of school dropout, then some students will receive the CCT even though they would have finished their primary or secondary education without any intervention. Conversely, other students who are at risk of leaving school will never have received the CCT. Both cases involve a problem of misidentification, and the consequence is an ineffective use of resources. However, there are relevant trade-offs involved in deciding what targeting mechanism to use. Targeting CCTs exclusively according to the likelihood of dropping out of school would weaken the program's ability to find the poor. This challenge is well addressed by Maluccio (2009), who also states that although "there certainly would be overlap among the beneficiary households selected under various possible approaches . . . , they almost certainly would not yield identical groups of beneficiaries" (p. 5).[4] My main contribution is to offer a methodology that improves targeting design and assessment when two or more dimensions or target groups matter and there are trade-offs between a range of potential targeting mechanisms. CCTs are just one example

---

reached by a program. The latter focuses on the (potential) effect of the program after implementation. The literature on the impact of CCTs on school enrollment is vast, especially in Latin America. For example, positive effects of CCTs on school enrollment have been found in Colombia (Attanasio and others, 2010; Barrera-Osorio and others, 2011), Ecuador (Schady and Araujo, 2008), Honduras (Galiani and McEwan, 2013; Glewwe and Olinto, 2004), Mexico (Schultz, 2004), and Nicaragua (Maluccio and Flores, 2005).

4. Not considering potential dropouts when targeting CCTs would be less of a cause for concern if school dropout were a negligible problem. But in Latin America the graduation rate (among those one year older than the school finishing age but younger than twenty-seven) reached only 0.54 in the late 2000s (Bassi, Busso, and Muñoz, 2015). Similarly, dismissing potential dropouts in CCT targeting would be less of a problem in contexts where there is a high degree of overlap between the latter group and those living in poverty. However, this is not guaranteed. For example, in Chile in 2013, only 16.1 percent of young school dropouts (aged fifteen to nineteen years) lived in a poor household, while only 12.4 percent of poor adolescents had dropped out of school (Salas Opazo, Ormazabal, and Crespo, 2015).

where policymakers face these trade-offs. The methodology I introduce processes the trade-offs involved in using different targeting mechanisms into a single metric, which facilitates the comparison of alternative targeting mechanisms. My paper offers two indicators that combine information from two or more target groups, namely, a head count index and a measure of social welfare to assess targeting. By providing a foundation for improved targeting assessment, targeting design can be enhanced.

The paper uses the two-dimensional targeting challenge already set out for CCTs to analyze whether a proxy means test can effectively identify both the poor and future school dropouts and to assess this capacity relative to alternative targeting mechanisms. I use rich administrative data sets from Chile to simulate different targeting mechanisms, with a focus on a predictive model of school dropout derived using a range of machine learning algorithms (MLAs), one of their first applications for school dropout outside a developed country. I then assess the targeting effectiveness of the PMT, the predictive model, and mechanisms combining both sources of information.

In my targeting assessment, there is a trade-off between using the PMT relative to using the MLA-based predictive model. When I use the PMT to target a hypothetical CCT, the targeting indicators associated with the poor improve, but the indicators related to dropouts worsen.[5] The opposite also holds. For different fixed budgets, total leakage (that is, the fraction of students receiving the CCT who are neither poor nor future dropouts) is minimized when I use both instruments in conjunction with each other. In other words, it is more effective to combine the predictive model and the PMT than to use them independently. However, this is not true when the social valuation of the two target groups differs to a large extent. If allocating the CCT to a poor student is four times more valuable than allocating it to a future dropout, or vice versa, the optimal approach is to use solely the mechanism designed to find the target group that is valued the most. These

5. Ideally, the metric to be used to assess targeting should be guided by the explicit goals of a program. In this case, the paper evaluates a hypothetical CCT program in which both the poor and future school dropouts are key target groups. I do not assess targeting for a specific Chilean CCT for two reasons. First, the hypothetical approach provides a broader perspective, going well beyond the country's specific case. Second, in Chile, unlike in other countries, multiple social policies could be labeled CCTs. These have comparable designs but target different populations. For example, *Asignación Social* (later *Ingreso Ético Familiar)* provided cash transfers to the poor conditional on the children and adolescents being enrolled in school (Universidad del Desarrollo, 2014). In contrast, *Beca de Apoyo a la Retención Escolar* targeted students with the highest risk of dropping out of school, whether poor or not, and provided cash transfers conditional on not dropping out of school (Salas Opazo, Ormazabal, and Crespo, 2015).

results point to the merit of using other targeting mechanisms instead of PMTs for CCTs: in contexts where public officials value finding the poor and future school dropouts equally, targeting can be improved when other dimensions beyond income are considered in the design.

The literature contains few attempts to assess CCT targeting that consider more dimensions than just income. A notable exception is Azevedo and Robles (2013), who assess the targeting performance of a CCT in Mexico using multiple indicators for each dimension. They find that their multidimensional targeting approach is better suited to identifying beneficiaries with higher rates of school nonattendance and child labor. This is comparable to my results. However, they also find that their model identifies the income poor as well as the mechanism used by the CCT. Since there is no trade-off between the two targeting mechanisms assessed, using multidimensional targeting is superior.

The latter is a key difference relative to my paper. The specific scenario of a single alternative that optimizes targeting for all relevant dimensions is unlikely to hold in every situation. If each of the available targeting mechanisms is more predictive of a specific target group, the choice of one mechanism over the other(s) will likely involve trade-offs. Opting for multidimensional targeting will not make these trade-offs disappear. In this scenario, the assessment approach advanced by Azevedo and Robles (2013) would not provide clear guidance for policymakers. Since a given targeting mechanism would reach a specific group but not the other(s), it would not be clear which targeting mechanism was optimal. To facilitate this decision-making process, my paper emphasizes indicators that combine information from the target groups, which, unlike unidimensional metrics, eases the assessment and processing of targeting mechanisms' trade-offs. Thus, while both papers offer a multidimensional targeting approach when multiple groups are relevant, Azevedo and Robles's approach is prescriptive for policymaking only in the absence of trade-offs between potential targeting mechanisms and the related target groups. My paper relaxes this restriction. Therefore, my proposed methodology is helpful for policymakers who need to target multiple groups, as it provides a single metric for comparing potential targeting mechanisms.

The two approaches also differ in a number of other ways.[6] Azevedo and Robles (2013) use few indicators to identify deprivation or risk in the

6. Their paper focuses on three dimensions (income, health, and education) at the household level, while I focus on two dimensions (income and education) for individuals in a specific age range for which these dimensions are critical.

educational dimension. Additionally, they use normative criteria (selecting specific indicators, thresholds, and weights) for this purpose. My approach uses a larger pool of variables, which allows for predicting empirically which adolescents will drop out of school, and thus does not need to use thresholds to define deprivations. Overall, my approach builds on data-driven models that can be more efficient in a rich data context. As more administrative data become available for public officers, the importance of predictive modeling increases for policy targeting.

The following features summarize the general framework used. The methodology intends to solve the problem faced by a policymaker who is interested in targeting a policy to multiple groups. This results in trade-offs between a range of available targeting mechanisms, such as predictive models, whereby using a given targeting mechanism improves the identification of a specific target group over the other(s). The policymaker seeks to optimize a single measure—namely, leakage or welfare—that incorporates information on all the relevant target groups. Using one metric facilitates processing the trade-offs associated with each of the available options. Depending on the metric chosen, the optimal targeting approach will derive from a single targeting mechanism or a combination thereof. Simulations are required to find this optimal approach. This general framework can be tailored to work using any number of target groups and available targeting mechanisms.

Overall, the paper offers a broad perspective on the challenge of multidimensional targeting and assessment, beyond CCT programs. The results also contribute to enriching the theoretical literature that seeks to minimize poverty or maximize social welfare (Coady and Skoufias, 2004; De Wachter and Galiani, 2006; Glewwe, 1992; Ravallion and Chao, 1989). As in the case of CCT allocation design and evaluation, moving from considering only one dimension in these theoretical models toward considering multiple dimensions seems desirable. For example, in welfare maximization models, it might be necessary to consider the utility provided not only by the transfer through the income dimension but also by the prevention of future dropout (and other relevant outcomes). In other cases, it might be necessary to include the elasticity of school dropout to extra income. Finally, regarding poverty minimization problems, it may be useful to incorporate future poverty alleviation explained by increased schooling in addition to current poverty alleviation due to the transfer.

In summary, the paper provides novel contributions to policy targeting. Overall, the findings are relevant not only for the specific Chilean CCT case but for all countries that wish either to develop predictive models using administrative records or to strengthen the targeting of their policies when multiple target groups matter.

The paper unfolds as follows. The next section introduces the data and describes the methods I use in developing the predictive model of school dropout and the targeting assessment. The paper then presents the results of the MLA predicting school dropout before showing the findings for the targeting assessment. The concluding section summarizes the paper's findings and comments on its contributions and implications.

## Data and Methods

This section describes in detail the data and methods. The first subsection introduces the data. I then present the methodological approach of the predictive model of school dropout and elaborate on the procedures and indicators of the targeting assessment. Finally, I explain how the data set is structured for the analysis.

### *Data*

Most of the data sets used in this analysis were provided by the Ministry of Social Development. I requested the data sets using the public procedures established by the Chilean government (Freedom of Information requests). I combine the data sets using individual identification numbers, which for privacy purposes were changed by the Ministry of Social Development using an algorithm that is unknown to me. The two most important sources of information in this research are the Ministry of Education performance data set and the social protection file (SPF) data set.

**MINISTRY OF EDUCATION PERFORMANCE DATA SET.** This data set contains information for the entire population of students who finish an academic year in primary and secondary education, excluding only students in differential education and flexible adult education. Each yearly data set has approximately 2,950,000 observations (one per student). I requested eight data sets (from 2009 through 2016) for this paper.

Some of the variables available in this data set include the following: school identification number (9,500 unique values), type of school (with categories such as traditional primary education and scientific-humanistic or technical-professional secondary education), grade (first through twelfth), academic performance, percentage of attendance, academic end-of-year classification, and student identification number. With this information, I created the following variables: school dropout (explained in the next subsection), school size, relative academic performance, relative attendance, school mobility, historical dropout rates by school, and academic cohort size.[7]

More educational information at the school level is available from public sources. Using the school identification number as a merge key, I obtained the schools' administrative status (such as public or subsidized private), geographic location (region), urban or rural status, average performance in language and mathematics on the national standardized test (SIMCE) for the grades that are most relevant to my sample (eighth and tenth), and management indicators.

**SOCIAL PROTECTION FILE DATA SET.** This data set contains information on Chilean households and all their members. It has a two-level structure. Each observation represents an individual (adult or child) who lives in a household. No individual can belong to more than one household. Each household has a unique identification number that allows for identifying all the individuals who belong to it.

Having an SPF is required to be eligible for multiple social policies, so households voluntarily request their SPF at the local government level. In January 2010, the data set had 10,782,270 individuals (Comité de Expertos de la Ficha de Protección Social, 2010), or approximately 63.5 percent of Chile's population. I use four of these data sets (from 2011 to 2014) in this research.

SPF scores estimate household income using variables correlated to household members' income. Therefore, the instrument is a form of a proxy means test. The household's score is mostly explained by the sum of the predicted labor income of each member.[8] Most of the variables that go into the formula are collected during a household interview. In theory, the SPF index ranks households from the poorest to the richest, similar to a ranking

7. Three variables define an academic cohort: the school, the type of education received within that school (for example, traditional or adult education; scientific-humanistic or technical-professional), and the grade in which the students were enrolled. Students belonging to the same cohort have these characteristics in common. Most schools have a specific orientation. However, some schools offer more than one type of education in a given grade (especially in secondary education). Students can also change streams from one academic year to the other.

8. This prediction is estimated for each household member of working age (in Chile, this is eighteen to sixty-four years for men and eighteen to fifty-nine years for women). The equation used for each household member depends on the characteristics of the individual. For those employed, variables regarding the features of their occupation are considered in the model. For each member, characteristics such as years of schooling are relevant in the prediction.

by per capita income. The SPF scale ranges from 2,072 points (the poorest households) to infinity (the richest) in theory.

Some of the SPF variables I access are the following: income, date of birth, proxy means test score (SPF score), gender, race, head of household, schooling, and employment. With this information, I can generate additional variables for each individual, such as poverty status (explained in detail in the third subsection) and the number of children under six years old in the household.

I combine the information from the social protection file with the Ministry of Education performance data set (at the individual level) to build variables for each academic cohort of students. These include average household income per capita, average schooling of the head of the household, and the share of students with a proxy means test score.

### *Methods: Predictive Model of School Dropout*

This subsection describes the methodological approach I take to build the predictive model of school dropout, including the predictors, the outcome, the characteristics of the prediction functions, and the criterion used to assess the predictions. In general terms, the problem I address in this part of the paper is to find the best function to predict future school dropout given the available information from the past. More formally,

$$Y_{it+k} = f\left(\mathbf{X}'_{it}, \mathbf{X}'_{it-1}, \ldots\ldots, \mathbf{X}'_{it-j}, \mathbf{Z}'_{i}\right),$$

I need to find a function $f$ that—given the vectors of variables $\mathbf{X}'_t$ (where $t$ is the year), $\mathbf{X}'_{t-1}, \ldots\ldots, \mathbf{X}'_{t-j}$, and $\mathbf{Z}'$ available for each individual $i$—produces, on average, the most accurate prediction of the outcome $Y$ in $t + k$. Given that the outcome, school dropout, is a dichotomous variable, this is a statistical classification problem, and $f$ is known as a classifier.

**THE PREDICTORS.** I include two types of predictors in the model. The first are contained in vectors $\mathbf{X}'_t, \mathbf{X}'_{t-1}, \ldots\ldots, \mathbf{X}'_{t-j}$. Specifically, $\mathbf{X}'_t$ is a vector of variables that change through time for student $i$ (such as academic performance, grade repetition, attendance, and mobility). The second group of predictors is embedded in $\mathbf{Z}'$, a vector of variables for student $i$ that do not vary through time (such as race) or that have only one observation (such as age).

The selection of variables included in the model is motivated by the literature on determinants of school dropout and bounded by the availability of administrative records. Rumberger and Lim (2008) summarize 203 studies for the United States over twenty-five years to identify statistically significant predictors of school dropout. Some individual characteristics that are relevant predictors are educational performance (for example, academic achievement, mobility, grade promotion, age, and the difference between age and expected age for the grade), behaviors (such as absenteeism, deviance, and employment), attitudes (like goals and self-perceptions), and background (for example, demographics and health). Their review also identifies institutional characteristics of students' families, schools, and communities. For example, the structure, practices, and financial and human resources of students' families are singled out as predictors. At the school level, they highlight the student composition, structural characteristics, resources, processes, and practices.

Hunt (2008) reviews the literature on factors associated with school dropout in developing countries. She identifies similar predictors to Rumberger and Lim (2008), but also adds other intrinsic challenges that these nations face, such as migration, conflict, and limited school supply.

The complete list of predictors is available in Crespo (2019).[9] There are fifty variables in total, which aim to cover all the dimensions highlighted by Rumberger and Lim (2008). As a result of the nature of the sources, the information is richer on educational performance and the characteristics of students' families than on predictors such as students' attitudes toward education.

Finally, Lamote and others (2013) argue that predictive models of school dropout need to account for the longitudinal and hierarchical structure of the data sets. This makes perfect sense due to the relevance of educational performance, which is a time-variant variable, and of schools and communities as predictors of future dropout. Accordingly, where feasible, all my models use three years of historical information ($\mathbf{X}'_t$, $\mathbf{X}'_{t-1}$, $\mathbf{X}'_{t-2}$) and include many variables that are at a higher level than the students.[10] The data set is appropriate for the task as it includes multiple strong predictors of school

9. The working paper version of this article contains a series of appendixes with supplementary material (available online at http://eprints.lse.ac.uk/101013/1/05_19_Cristian_Crespo.pdf).

10. There is a trade-off concerning how many years of historical information to use. As a result of how the data set is structured, adding a year can improve the prediction of school dropout, but it reduces the sample size (or, alternatively, leaves at least an entire cohort with no information for at least one year). I decided to use three years and only cohorts that have three years of historical information. Variables on $t - 2$ have some, albeit limited, predictive power (as shown in the next section), and this decision allowed me to pool four different cohorts that have all three years of historical information.

dropout and multiple years of information on academic attainment, mobility, and attendance, as well as information at the household level (such as years of schooling of its members and per capita income) and the school and academic-cohort levels.

**THE OUTCOME.** I use the Ministry of Education performance data set to identify students who dropped out of school. The process involves merging different years of the data set and linking observations by the student identification number. More precisely, I link each student in primary and secondary education who concluded their academic year $t$ and did not graduate from their secondary studies with him- or herself in years $t + 1$ and/or $t + 2$. Using this procedure, I identify students who dropped out of school after year $t$.[11] Student dropout can be measured in multiple ways. I use three different measures of school dropout to verify the consistency of the results:

—dropout_t1: The student finished the academic year $t$ and then failed to enroll in $t + 1$ or enrolled but withdrew before the end of year $t + 1$.

—dropout_t2: The student finished the academic year $t$ and (disregarding what happened in $t + 1$) then failed to enroll in $t + 2$ or enrolled but withdrew before the end of year $t + 2$.

—dropout_t12: The student finished the academic year $t$ and then failed to enroll in $t + 1$ or enrolled but withdrew before the end of year $t + 1$ or failed to enroll in $t + 2$ or enrolled but withdrew before the end of $t + 2$.

Insofar as students who drop out in $t + 1$ may or may not return to school in $t + 2$, a given student could be a dropout both years, thereby appearing in both dropout_t1 and dropout_t2. The variable dropout_t12 takes a value of one if, for a given student, either dropout_t1 or dropout_t2 (or both) takes a value of one. Thus, dropout_t12 can be interpreted as dropping out of school at any point within two years of completing an academic year.

**THE CLASSIFIER AND MACHINE LEARNING ALGORITHMS.** I determine $f$ using supervised machine learning algorithms. MLAs are a powerful and flexible provider of quality predictions and a helpful tool for prediction policy problems (Kleinberg and others, 2015; Mullainathan and Spiess, 2017).[12] MLAs have been used in research on such topics as recidivism, teacher hiring, and

11. Because of how I measure school dropout, the sample includes only students who finished academic year $t$.

12. The machine learning literature focuses mainly on the problem of prediction and not on capturing the relationship between the predictors and the outcome. Initially, MLAs were not designed to obtain deep structural parameters or causal inference (Nichols, 2018). However, an emerging literature connects MLAs with causal inference for policy (Abadie and others, 2014; Athey and Imbens, 2015a, 2015b).

identification of vulnerable groups. They find functions that predict well out of sample or do not overfit the data; they can discover a complex structure that is not specified in advance; and they allow researchers to manage high-dimensional settings in which the number of variables is larger than the number of observations (James and others, 2013; Mullainathan and Spiess, 2017).

MLAs are suitable in my case for three reasons. First, in theory, an approach that maximizes the predictions of an outcome outside the sample is preferred for a prediction policy problem (such as determining which students will drop out) relative to an approach that maximizes predictions within the sample. Second, a priori I ignore the structure of the function (for example, the number of variables to include) or the form that achieves the best prediction of school dropout. Using MLAs expands the likelihood of finding the best model because some MLAs consider interactions and polynomials while others directly address the challenge of variable selection. Finally, with machine learning I can better manage the number of parameters to include in the data set. Although I do not face a high-dimensionality problem, reducing the number of predictors (by not directly including higher-order terms) facilitates the calculations.

To obtain predictions that work well out of sample, machine learning uses a training data set and a test data set. The models must be estimated in the former data set and assessed with the latter. MLAs aim to avoid overfitting; in other words, they seek to optimize their predictions in the test data set (out of sample) rather than in the training data set (in sample). To do so, each algorithm first tries to determine its optimal level of complexity in the training data set. The specific indicators of model complexity vary by algorithm, but in general terms these are called regularizers. The less regularization there is, the better the in-sample predictions (Mullainathan and Spiess, 2017). These model-complexity parameters can be viewed as variables that can be tuned to produce optimal predictions in the test data set (Varian, 2014).

The last process is known as empirical tuning. It consists of fitting the algorithm in one part of the training data set and then determining the optimal value of the regularizer by assessing its prediction performance in another part of the training data set (Mullainathan and Spiess, 2017). Van der Vaart, Dudoit, and Van der Laan (2006) show that the effectiveness of the procedure is increased if the training data set is subdivided into multiple subsamples or folds. This is known as cross-validation, with five or ten folds being the most common (Mullainathan and Spiess, 2017). In this type of cross-validation, the regularizer with the best average performance is chosen.

MLAs vary in terms of their flexibility for finding the best *f*. Shrinkage methods such as lasso and elastic nets are the most restrictive because they can only generate linear functions (that is, with no interaction between the predictors or other higher-order terms). These algorithms are less flexible than ordinary least squares as there is a penalty for every regression coefficient that is different from zero, which leads to the coefficients of the linear regression being shrunk toward zero relative to least squares (James and others, 2013). Generalized additive models (GAMs) expand the range of shapes to estimate *f* from linear to more complex approaches, for example, some nonlinear relationships (James and others, 2013). In practice, GAMs fit a nonlinear function separately for each predictor and then add all these functions. Since the model is additive, interactions between the predictors are not considered.

Tree-based approaches admit interactions by stratifying the predictor space into regions (McBride and Nichols, 2016). For example, if only two predictors of school dropout are available (age and attendance), a classification tree algorithm can be as follows: a dropout is predicted only if a student is older than seventeen years and has an attendance of lower than 70 percent. Methods such as random forest and boosting are a combination of multiple trees.

Finally, a highly flexible approach uses support vector machines. In a classification problem, this algorithm aims to find a hyperplane separating the two classes. If this hyperplane cannot be found, a kernel trick is applied (Theodoridis and Koutroumbas, 2009). The feature space of the problem is expanded, and a new hyperplane is fitted in this transformed space. This process may produce nonlinear class boundaries in the original predictors' space.

James and others (2013) claim that no single algorithm is superior in every possible context. For this reason, I tried multiple MLAs. For simplicity, the paper presents results for only six of them: elastic nets (glmnet), GAMs, gradient boosting models (GBM), lasso, support vector machines (SVMs), and random forest (RF). These six MLAs use the same inputs, which are the fifty predictors described in Crespo (2019). I implement the MLAs in the R software using the caret package. Kuhn (2008) is a precious source for this purpose. I use ten-fold cross-validation and two test data sets, one to conduct out-of-sample validation and one to assess the quality of the predictions over time. The design of the training data set and the two test data sets are explained in the next subsection. With regard to the treatment of the predictors, I convert categorical variables into dummy variables, and the caret package carries out standardization on all the predictors before executing each MLA.

**TABLE 1. Confusion Matrix of a Classifier of Student Dropout**

| | *Predicted class* | | |
|---|---|---|---|
| *True class* | *Not a dropout* | *Dropout* | *Total* |
| Not a dropout | True negatives | False positives | Nondropouts |
| Dropout | False negatives | True positives | Dropouts |
| Total | Negatives | Positives | All population |

**TABLE 2. Indicators Used in the Predictive Model of School Dropout**

| *Name* | *Formula* |
|---|---|
| True positive rate or sensitivity | True positives/Dropouts |
| False positive rate or 1 – specificity | False positives/Nondropouts |
| Accuracy | (True negatives + True positives)/Total |

**THE CRITERION USED TO SELECT THE BEST *F*.** Statistical classification problems have only four possible outcomes for dropout prediction: a model either correctly predicts a dropout, or incorrectly predicts a dropout, or fails to predict a dropout, or correctly predicts a nondropout. More generally, these processes result in four categories, which are labeled true positives, false positives, false negatives, and true negatives. This can be summed up in a confusion matrix like the one in table 1. Multiple indicators derived from combinations of a confusion table have been used to report the quality of predictions. Within studies on dropout prediction, there is no standard metric that facilitates comparisons (Bowers, Sprott, and Taff, 2013). Following these authors, I provide true positive rates, false positive rates, and accuracy (table 2). An exhaustive list of this family of indicators is available in Crespo (2019).

A perfect classifier would achieve a true positive rate of one, with all dropouts predicted as such, and a false positive rate of zero, or no incorrect predictions of dropouts. No classifier achieves this performance. In practice, dropout prediction models tend to maximize the true positive rate (or sensitivity) and minimize the false positive rate (or 1 – specificity). Nonetheless, there is a trade-off between these two indicators. As a predictive model classifies more observations as dropouts, both the true positive rate and the false positive rate increase.

Receiver operating characteristics (ROC) curves summarize this trade-off. An ROC curve simultaneously displays the false positive rate (horizontal axis) and the true positive rate (vertical axis) given by a classifier, representing all possible outputs or scenarios (James and others, 2013). Thus the area under the curve (AUC) provides a measure of the overall predictive performance of the classifier. The AUC scale ranges from zero to one: the better the classifier is, the closer its AUC will be to one. Conversely, a classifier making predictions at random has an expected AUC of 0.5.

The AUC is a useful indicator for comparing the overall performance of multiple predictive models. Models with a higher AUC are, on average, better at statistical classification relative to models with a lower AUC. A model with an ROC curve that is on top of other curves all the way along the horizontal axis is unambiguously a better classifier in every possible scenario.

I use the AUC estimates to select the best *f*. I calculate these in the first test data set (for out-of-sample validation) and for the three measures of school dropout introduced at the beginning of this subsection. The advantages of using the AUC are twofold. In the first instance, the performance of the MLA predicting school dropout can be compared graphically. Second, the AUC integrates in one value all the potential classification outputs of each algorithm. This feature frees me to select an arbitrary threshold to assess the performance of the classifiers (such as choosing the algorithm with the highest true positive rate when the false positive rate reaches 0.20).

Additionally, I provide true positive rates, false positive rates, and accuracy for two specific scenarios. I force the MLA to classify 10 percent and 30 percent of students as future dropouts. These indicators help to establish comparisons with the outputs obtained by other scholars.

### *Methods: Targeting Assessment*

After identifying the best-performing algorithm, I can use two indexes to target a hypothetical CCT. The first is the proxy means test score from the social protection file. I derive the second from the outputs of the best MLA ($f$). Each of these outputs represents the probability that the model is observing a future school dropout.

This subsection describes the methods related to the targeting assessment of a hypothetical CCT on the poor and on future school dropouts. After explaining how I construct the poverty variable, I elaborate on the indicators used to assess targeting, namely, total leakage and leaked welfare. The subsection closes with a discussion of policy alternatives.

**POVERTY.** Poverty status is not directly available from the SPF data set. However, it is possible to build an estimate of poverty status using household structure and income (in the SPF, the most relevant sources of income are labor and pensions). There are many approaches to constructing this variable.

I use total household income over number of members. Using per capita income is consistent with the traditional methodology used in Chile to measure poverty. I define a student as poor if he or she is part of the lowest quintile of per capita income in the sample. I chose this poverty line because the poverty rate was approximately 20 percent in the population analyzed in one year of the assessment.[13]

**TOTAL LEAKAGE.** The literature on poverty targeting offers multiple indicators of targeting effectiveness. One example is the AUC in ROC analysis discussed above (Baulch, 2002; Wodon, 1997). Another common approach is to provide undercoverage and leakage rates (Coady, Grosh, and Hoddinott, 2004). The undercoverage rate is the proportion of poor households or individuals not receiving the program. The leakage rate is the fraction of nonpoor among those who are receiving the program. The use of these rates has two common limitations (Coady and Skoufias, 2004). First, they disregard distributional information; for example, giving a transfer to someone in the highest 1 percent of income counts the same as giving it to someone marginally over the poverty line. Second, the size of the transfer is irrelevant; that is, it does not make a difference whether a poor household receives a minuscule transfer or an amount that lifts it over the poverty line. One of the preferred ways to address this latter limitation is to assess targeting based on the impact on poverty (Grosh and Baker, 1995; Skoufias, Davis, and de la Vega, 2001).

Although using leakage and undercoverage rates restricts the depth of the analysis on the poverty dimension, it facilitates the comparisons of targeting indicators for poverty and school dropout. It also facilitates combining future school dropouts and the poor into one indicator. Thus, despite their limitations, I opt to use these types of indicators to assess the performance of targeting mechanisms. Five of the indicators I use in the paper are presented in table 3.

Total leakage, defined as the share of nondropouts and nonpoor receiving the CCT after the simulation, can be interpreted as the inclusion error (see table 4). Students (potential recipients of a CCT) fall into one of four classes. Either they are poor and will drop out of school, or they are poor but will not drop out, or they are not poor but will drop out, or they are not poor and will not drop out. Targeting is unsuccessful when a CCT is given to the fourth type of student because no target group is reached.

13. This poverty rate is higher than the official poverty rate. However, my sample is not representative of the whole student body. Using the Chilean CASEN survey, I estimate the poverty rate in 2011 for the population of students who are most likely to constitute my sample. I obtain an estimate of 20.06 percent using the traditional methodology.

**TABLE 3. Indicators Used in the Targeting Assessment**

| Name | Formula |
|---|---|
| Poor undercoverage | No. poor not receiving CCT/No. poor |
| Nonpoor leakage | No. nonpoor receiving CCT/No. receiving CCT |
| Dropout undercoverage | No. future dropouts not receiving CCT/No. future dropouts |
| Nondropout leakage | No. nondropouts receiving CCT/No. receiving CCT |
| Total leakage | No. nondropouts and nonpoor receiving CCT/No. receiving CCT |

**TABLE 4. Successful Targeting and Targeting Errors in the Context of Two Target Groups**

| | Hypothetical CCT recipient | |
|---|---|---|
| True class | No | Yes |
| Nonpoor and nondropout | Successful targeting | Inclusion error |
| Nonpoor and dropout | Exclusion error | Successful targeting |
| Poor and nondropout | Exclusion error | Successful targeting |
| Poor and dropout | Exclusion error | Successful targeting |

The selection of total leakage as the first main indicator of my analysis is justified on theoretical grounds. One minus leakage can be equivalent to the distributional characteristic (DC), a cost-benefit statistic used to compare the welfare impact of transfers with a common budget (Coady and Skoufias, 2004). The authors show that the DC, $\lambda$, for any given scheme $j$ is

$$\lambda_j = \sum_h \beta^h \theta^h,$$

where $\beta^h$ is the social valuation (welfare weight) of extra income to household $h$, and $\theta^h$ represents the share of the total program budget received by household $h$.

An advantage of the DC is that welfare weights are made explicit, and it generalizes from simpler to more complex cases (Coady, Grosh, and Hoddinott, 2004). When the size of the transfer is identical for each household and the social valuation of extra income is equal to one for a poor household and zero otherwise, the DC indicator is equivalent to one minus the leakage rate:

$$\lambda_j = \sum_h \beta^h \theta = \theta \sum_h \beta^h = \frac{\sum_h \beta^h}{\text{No. recipients}} = \frac{\text{No. poor recipients}}{\text{No. recipients}}$$

$$= 1 - \frac{\text{No. nonpoor recipients}}{\text{No. recipients}} = 1 - \text{Leakage.}$$

Under some additional assumptions, when the size of a CCT is identical for each individual and when the social valuation of income is equal to one for any CCT recipient who is either poor or a future dropout and zero otherwise, the DC indicator is equivalent to one minus total leakage.

Total leakage is the cornerstone indicator that I use in my research to compare the targeting performance of alternative instruments. The indicator has the major advantage of allowing for the integration of two important target groups for CCTs. Additionally, the logic behind this indicator is useful for other parts of the assessment, when I focus on social welfare and targeting costs.

Peyre Dutrey (2007) criticizes the use of leakage in targeting assessments because it does not account for individuals who are excluded—that is, it does not consider undercoverage. However, leakage and undercoverage rates are related. If coverage increases (and undercoverage decreases), leakage is likely to increase. Therefore, rather than seeking the optimal rate of undercoverage and leakage, I assess three different coverage levels, or budget allocations, of a hypothetical CCT. I explain this aspect of the paper in detail at the end of this subsection. Overall, within a fixed budget and coverage rate, the targeting mechanism with the lowest total leakage is optimal.

**LEAKED WELFARE.** I also analyze whether the findings of the targeting assessment hold when I change the social valuation of the target groups. Up to this point, I have implicitly assumed that successfully targeting a student who is poor is as socially worthwhile as correctly targeting a student who will drop out of school. I introduce four different scenarios of social valuation across the two target groups. In the first two scenarios, each target group is twice as important as the other; in the last two scenarios, the difference in valuation increases to four times the other target group. The choice of these scenarios does not have any theoretical justification but rather is merely practical. Following the logic of the DC, the welfare impact of a transfer scheme $j$, which provides an equal amount for each individual $i$, can be measured by the following formula:

$$\lambda_j = \omega \sum_i \gamma^i = \frac{1}{\text{No. recipients}} \sum_i \gamma^i = \frac{\sum_i \gamma^i}{\text{No. recipients}},$$

where $\omega$ is the share of the total program budget received by each adolescent who is a CCT recipient and $\gamma^i$ is the social valuation (or welfare weight) of extra income to adolescent $i$.

**TABLE 5. Social Valuation in Different Scenarios**

| True class | Social valuation scenarios: The poor are twice as important as dropouts | The poor are four times more important than dropouts | Dropouts are twice as important as the poor | Dropouts are four times more important than the poor |
|---|---|---|---|---|
| Nonpoor and nondropout | 0 | 0 | 0 | 0 |
| Nonpoor and dropout | 1/3 | 1/5 | 2/3 | 4/5 |
| Poor and nondropout | 2/3 | 4/5 | 1/3 | 1/5 |
| Poor and dropout | 1 | 1 | 1 | 1 |

Note: Social valuation is the welfare weight, $\gamma$.

For $\lambda_j$ to have minimum and maximum values of zero and one, I choose the welfare weights using the following logic. A hypothetical CCT recipient $i$ who is neither a future dropout nor poor receives a $\gamma$ value of zero. Conversely, each CCT recipient $i$ who belongs to both target groups receives a $\gamma$ value of one. The social valuations of each class of student in each of the four scenarios I use in the paper are presented in table 5.

The targeting mechanism $j$ that provides the highest $\lambda j$ maximizes welfare. Given the weights I use, the last statement can be rephrased as follows: the targeting mechanism $j$ that provides the lowest $1 - \lambda_j$ maximizes welfare (for any given budget). This last indicator is the focus of the welfare assessment. For simplicity, I refer to it as leaked welfare. More formally,

$$\text{Leaked welfare} = 1 - \lambda_j = 1 - \frac{\sum_i \gamma^i}{\text{No. recipients}}.$$

**POLICY ALTERNATIVES: BUDGET, COVERAGE, AND TARGETING MECHANISMS.** CCTs are not universal schemes. Stampini and Tornarolli (2012) show that coverage varies by year and country in Latin America. Consequently, I repeat my targeting assessment for different levels of coverage for a hypothetical CCT. Given that in my study the transfer size remains unchanged, an increase or decrease in the CCT program budget only affects coverage. For this reason, I repeat my targeting assessment for three different budget scenarios for a hypothetical CCT, assuming no administrative costs in the first instance. In the first case, the budget allows for reaching only 5 percent of the students in the sample. In the second and third scenarios, the budget allows for reaching 20 and 40 percent of the sample, respectively. These three cases aim to re-create real policy environments: a narrowly targeted CCT, a CCT whose coverage is in line with the population living in poverty, and a broadly targeted CCT.

I begin by looking at the targeting performance separately for each instrument. First, I use only the proxy means test score of the SPF. Second, I use the predictions derived from the best *f*. The assessment continues with two combined mechanisms. I target a hypothetical CCT assigning the first 25 percent of the available budget using the PMT score and the remaining 75 percent with the predictive model; I then reverse the percentages for the second mechanism. For example, when the budget allows for reaching 20 percent of the students in the sample with a CCT and I allocate this using the second combined approach (with 75 percent assigned first using the SPF), the procedure works as follows. I first select the 15 percent of the sample with the lowest PMT scores and assign them the CCT. I then choose the remaining 5 percent of the students by observing the highest likelihood of dropping out among those not selected in the first step.

### *Sample and Data Set Structure*

The sample excludes students below seventh grade in year $t$, younger than twelve years old by June of year $t$, and over twenty-one years old by March of year $t + 1$. I apply these restrictions considering that student dropout in Chile is a cause for concern mainly in secondary school and that twenty-one years old is the maximum age for enrolling in traditional secondary education.[14]

Another crucial characteristic of the sample is that it includes only adolescents in the SPF registry. Thus this is not a representative sample of the population, as high-income households were less likely to request an SPF. These features do not favor making inferences about the whole student body. However, this is not problematic if the findings are linked to a subset of the entire population: namely, students with an SPF. Insofar as this subset is more likely to include recipients of social programs, the findings of this study remain relevant. In practice, 26.3 percent of adolescents did not make it into my final sample because of the lack of an SPF.

To undertake the MLAs and the targeting assessment, I structure the data set based on four year-cohorts $t$ ($t$ = 2011, 2012, 2013, 2014), using

14. In the Chilean educational system, students are, in theory, expected to graduate from secondary education at the age of eighteen. However, grade repetition and school dropout can delay graduation from secondary studies.

**TABLE 6. Data Set Structure**

| *Academic cohort* | *Academic and school information* | *SPF info* | *Dropout information* |
|---|---|---|---|
| 2011 | 2009 ($t-2$), 2010 ($t-1$), and 2011 ($t$) | 2011 ($t$) | 2012 ($t+1$) and/or 2013 ($t+2$) |
| 2012 | 2010 ($t-2$), 2011 ($t-1$), and 2012 ($t$) | 2012 ($t$) | 2013 ($t+1$) and/or 2014 ($t+2$) |
| 2013 | 2011 ($t-2$), 2012 ($t-1$), and 2013 ($t$) | 2013 ($t$) | 2014 ($t+1$) and/or 2015 ($t+2$) |
| 2014 | 2012 ($t-2$), 2013 ($t-1$), and 2014 ($t$) | 2014 ($t$) | 2015 ($t+1$) and/or 2016 ($t+2$) |

Notes: Academic, school, and dropout information is from the Chilean Ministry of Education. SPF data are from the Chilean Ministry of Social Development.

information from $t-2$, $t-1$, $t$, $t+1$, and $t+2$ for each individual in the cohort. Hence, each cohort on its own is a panel data set. I pool these four cohorts to obtain the full data set, which thus contains observations from eight years (2009 through 2016) (see table 6).

I divide the full data set into two parts, termed old and new. The old subset contains cohorts 2011, 2012, and 2013. I partition this subset into a training data set and a test data set using random assignment. Each observation in the old subset has a 0.75 probability of ending up in the training data set, which is used to train the MLAs. I then test the algorithms and implement the targeting assessment in the test data set. The new subset, which contains the 2014 cohort, is only used to assess the quality of the predictions of school dropout over time. This process is called out-of-time validation, and the results are available in Crespo (2019).

## Results: Predictive Model of School Dropout

This section starts with a review of the summary statistics of school dropout and multiple variables included in the model. The second part focuses on the results of the MLA predicting school dropout. I provide ROC curves, their AUC, true positive rates, false positive rates, and accuracy for three measures of school dropout. I also analyze which variables of the model mostly explain the variation in school dropout.

### *Summary Statistics*

Table 7 provides summary statistics for some individual-level variables. The average dropout rates for years $t+1$ and $t+2$ are 0.06 and 0.09, respectively, for the 2011–13 period. Within this time range, eleven out of 100 adolescents dropped out in either year $t+1$ or $t+2$. All measures of dropout declined

**TABLE 7. Summary Statistics by Year-Cohort**

| | Year-Cohort | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *2011* | *2012* | *2013* | *2014* | *2011–13* | | | |
| *Variable* | *Mean* | | | | *Mean* | *Std. dev.* | *Min.* | *Max.* |
| *A. Dropout information* | | | | | | | | |
| Dropout in $t+1$ | 0.07 | 0.06 | 0.06 | 0.05 | 0.06 | 0.25 | 0 | 1 |
| Dropout in $t+2$ | 0.10 | 0.09 | 0.08 | 0.07 | 0.09 | 0.28 | 0 | 1 |
| Dropout in $t+1$ or $t+2$ | 0.12 | 0.11 | 0.10 | 0.09 | 0.11 | 0.31 | 0 | 1 |
| Dropout in $t+2$ (if dropout in $t+1=1$) | 0.67 | 0.64 | 0.63 | 0.62 | 0.65 | 0.48 | 0 | 1 |
| Dropout in $t+1$ (if dropout in $t+2=1$) | 0.49 | 0.48 | 0.47 | 0.47 | 0.48 | 0.50 | 0 | 1 |
| *B. Academic information* | | | | | | | | |
| Grade point average in $t$ | 5.27 | 5.31 | 5.32 | 5.36 | 5.30 | 0.68 | 1 | 7 |
| Attendance in $t$ (%) | 88.10 | 90.90 | 90.30 | 90.40 | 89.70 | 11.60 | 1 | 100 |
| Promoted in $t$ | 0.88 | 0.91 | 0.91 | 0.92 | 0.90 | 0.31 | 0 | 1 |
| Mobility in $t$ | 0.25 | 0.24 | 0.24 | 0.23 | 0.24 | 0.43 | 0 | 1 |
| Grade point average in $t-1$ | 5.41 | 5.37 | 5.40 | 5.41 | 5.39 | 0.63 | 1 | 7 |
| Attendance in $t-1$ (%) | 92.80 | 90.10 | 92.10 | 91.50 | 91.70 | 9.30 | 1 | 100 |
| Promoted in $t-1$ | 0.94 | 0.92 | 0.94 | 0.94 | 0.93 | 0.26 | 0 | 1 |
| Mobility in $t-1$ | 0.24 | 0.24 | 0.23 | 0.23 | 0.23 | 0.42 | 0 | 1 |
| *C. School information (year t)* | | | | | | | | |
| Primary traditional school | 0.40 | 0.41 | 0.41 | 0.41 | 0.40 | 0.49 | 0 | 1 |
| Secondary traditional SH school | 0.34 | 0.35 | 0.36 | 0.36 | 0.35 | 0.48 | 0 | 1 |
| Secondary traditional TP school | 0.23 | 0.21 | 0.20 | 0.20 | 0.21 | 0.41 | 0 | 1 |
| Public school | 0.48 | 0.46 | 0.45 | 0.44 | 0.46 | 0.50 | 0 | 1 |
| Private subsidized school | 0.48 | 0.50 | 0.50 | 0.51 | 0.49 | 0.50 | 0 | 1 |
| Number of students in the academic cohort | 117.10 | 110.80 | 108.30 | 106.10 | 112.10 | 103.20 | 1 | 1324 |
| Rural school | 0.08 | 0.08 | 0.07 | 0.07 | 0.07 | 0.26 | 0 | 1 |
| Dropout rate in previous academic cohort | 0.07 | 0.07 | 0.06 | 0.06 | 0.07 | 0.10 | 0 | 1 |
| School language score on SIMCE | 249.8 | 251.5 | 251.1 | 248.7 | 250.8 | 26.3 | 142 | 345 |
| School math score on SIMCE | 248.7 | 249.6 | 253.7 | 255.5 | 250.7 | 32.7 | 138 | 381 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *D. Social protection file information* | | | | | | | | |
| Age at the end of academic year $t$ (years) | 15.39 | 15.39 | 15.39 | 15.41 | 15.39 | 1.65 | 12.50 | 20.67 |
| Male | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0 | 1 |
| Indigenous background | 0.08 | 0.09 | 0.09 | 0.09 | 0.09 | 0.28 | 0 | 1 |
| Household number of rooms | 2.18 | 2.15 | 2.11 | 2.10 | 2.15 | 0.93 | 0 | 63 |
| Head of household (HH) lives with a partner | 0.60 | 0.59 | 0.58 | 0.57 | 0.59 | 0.49 | 0 | 1 |
| HH is female | 0.44 | 0.45 | 0.46 | 0.48 | 0.45 | 0.50 | 0 | 1 |
| HH years of schooling | 9.46 | 9.62 | 9.70 | 9.83 | 9.59 | 3.41 | 0 | 24 |
| HH employed and with social security | 0.41 | 0.41 | 0.41 | 0.41 | 0.41 | 0.49 | 0 | 1 |
| Household size | 4.30 | 4.24 | 4.24 | 4.21 | 4.26 | 1.45 | 2 | 34 |
| Real household per capita income (CLP) | 64,218.9 | 66,939.9 | 66,783.0 | 70,188.6 | 65,963.9 | 57,529.2 | 0 | 5,850,206 |
| Social protection file PMT score | 7,537.2 | 7,384.2 | 7,193.6 | 7,025.8 | 7,373.0 | 3,733.3 | 2,072 | 15,625 |
| No. observations (student level) | 960,514 | 930,225 | 937,843 | 968,584 | 2,828,582 | | | |
| No. observations (school level) | 7,143 | 7,135 | 7,117 | 7,097 | 7,354 | | | |

Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

Notes: I link each student in primary and secondary education who concluded their academic year $t$ and did not graduate from their secondary studies with him- or herself in years $t+1$ and/or $t+2$. Using this procedure, I identify the students who dropped out of school after year $t$ (in $t+1$, $t+2$, or either $t+1$ or $t+2$). Dropping out of school means failing to enroll or enrolling but withdrawing before the end of the academic year. The mobility rate in $t$ and $t-1$ corresponds to the proportion of students switching schools between $t-1$ and $t$, and $t-2$ and $t-1$ respectively. SH school, scientific-humanistic school; TP school, technical-professional school.

annually from 2011 to 2014: from 0.07 to 0.05 for adolescents dropping out in year $t + 1$, from 0.10 to 0.07 for adolescents dropping out in year $t + 2$, and from 0.12 to 0.09 for adolescents dropping out in either year $t + 1$ or year $t + 2$. With regard to the dynamics of dropout, for the 2011, 2012, and 2013 cohorts, sixty-five out of 100 adolescents who dropped out in year $t + 1$ did not return to school in year $t + 2$, on average. Among those who were dropouts in year $t + 2$, only forty-eight out of 100 adolescents had previously dropped out in year $t + 1$, while fifty-two out of 100 dropped out only in year $t + 2$.

Regarding academic information in year *t*, between 2011 and 2013, adolescents had a grade point average of 5.30; their attendance was 89.7 percent; nine out of ten students were promoted to the next grade; and their mobility rate in year *t* was 0.24. The grade point average and promotion rate of students increased marginally from 2011 to 2014.

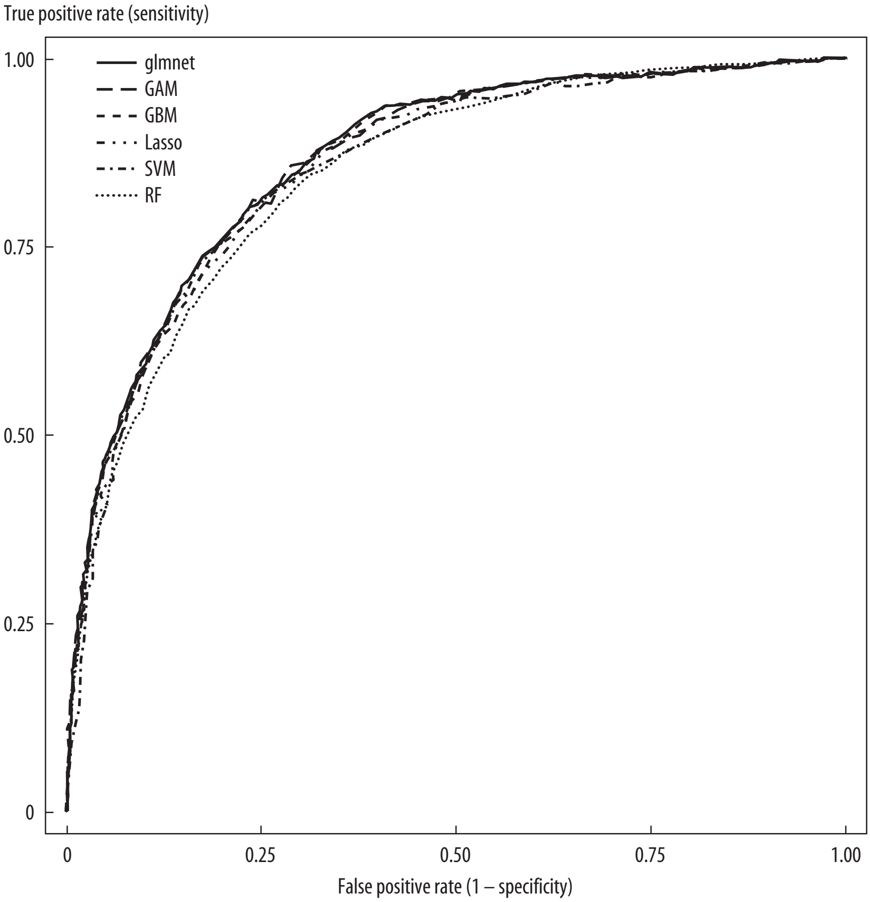
Between 2011 and 2013, four out of ten adolescents attended traditional primary education in year *t,* while 35 and 21 percent of adolescents were enlisted in traditional secondary education, in scientific-humanistic (SH) and technical-professional (TP) schools, respectively. In the period, 49 percent were enrolled in subsidized private schools and 46 percent in public schools.

Based on the SPF records, between 2011 and 2013, on average, adolescents were 15.39 years old by the end of the academic year *t*, half of the students were male, and nine out of 100 were indigenous. Concerning the heads of household, 45 percent were females, 59 percent lived with a partner, 41 percent were employed and contributing to social security, and their average schooling was 9.59 years. Between 2011 and 2013, student households had an average of 4.26 members, living in 2.15 rooms. The average monthly real income per capita was CLP 61,012 (equivalent to U.S. $116.50 at the December 30, 2013, exchange rate).

### *Results of Models Predicting School Dropout*

Figure 1 presents the ROC curves for six MLAs predicting dropout_t12. As the figure shows, the curves of the six models are close to each other, and no single one is above or below the rest along the whole horizontal axis. This suggests that the six MLAs have minor differences in terms of the area under the ROC curve. The elastic net algorithm (glmnet) curve has a higher degree of convexity and tends to be above all the other curves in a broad range of false positive rates. Conversely, the random forest curve is below the others in some sections of the graph (for example, where the true positive rate lies between 0.50 and 0.75).

**FIGURE 1. ROC Curve for Models Predicting School Dropout in Year $t+1$ or $t+2$**

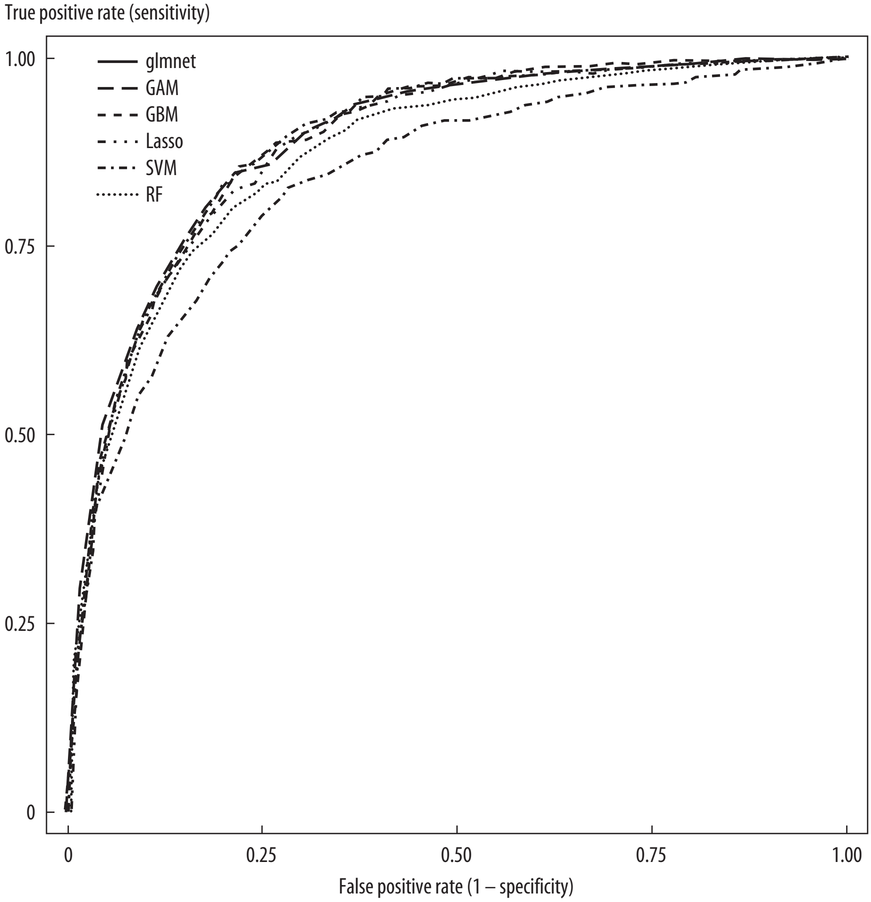



Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

Figures 2 and 3 graph the ROC curves predicting dropout_t1 and dropout_t2. The patterns are similar to figure 1. In both figures, the glmnet curves are predominantly above the other curves. However, the GBM and GAM curves closely follow and even surpass the glmnet along some parts of the horizontal axis. On the other hand, the SVM algorithm is unambiguously the worst performer in these assessments.

Table 8 presents the area under the ROC curve for each of the six MLAs. The glmnet has the largest AUC for all three measures of dropout: 0.866 for

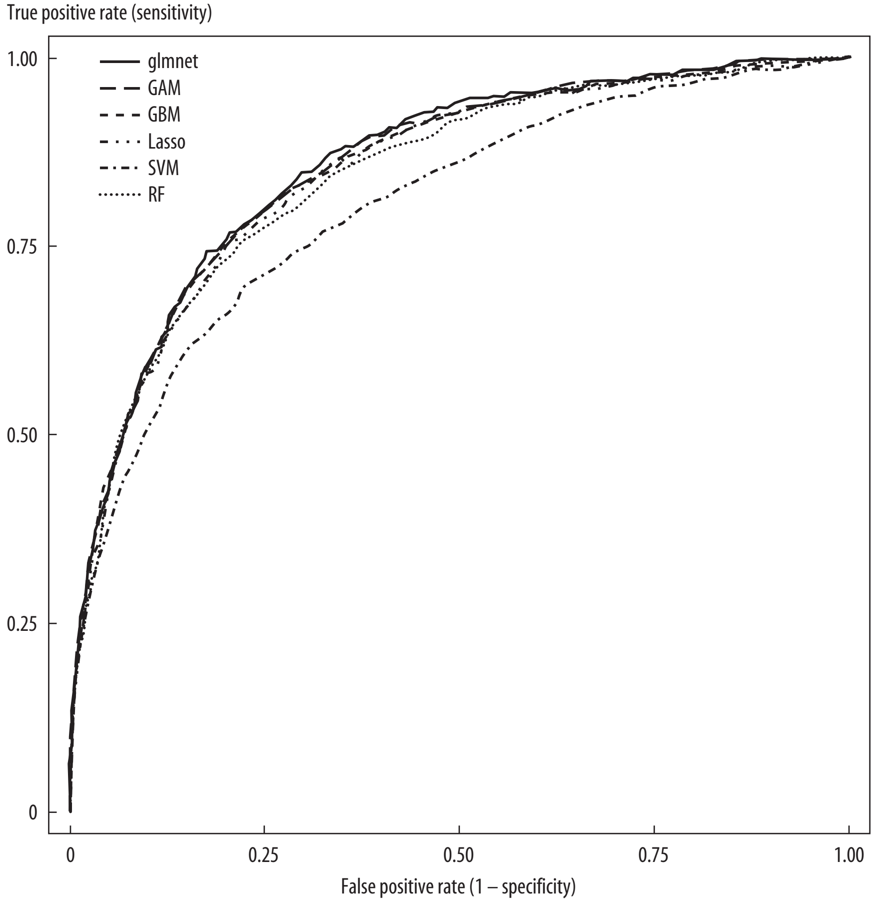
**FIGURE 2. ROC Curve for Models Predicting School Dropout in Year $t+1$**




Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

dropout_t12, 0.893 for dropout_t1, and 0.857 for dropout_t2. The GAM models have the second-highest AUC in all three cases, followed by the GBM models. More generally, for these three algorithms, the AUC is above 0.860 in the classification of dropout within two years, 0.890 in the classification of dropout after one year, and over 0.850 in the second-year dropout classification. Conversely, RF and SVM algorithms have the worst performances on all three measures of school dropout.

**FIGURE 3. ROC Curve for Models Predicting School Dropout in Year *t* + 2**




Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

**TABLE 8. Area under the ROC Curve for Models Predicting School Dropout**

| *Machine learning algorithm* | *School dropout measures* | | |
|---|---|---|---|
| | *Dropout in t + 1 or t + 2* | *Dropout in t + 1* | *Dropout in t + 2* |
| glmnet | 0.866 | 0.893 | 0.857 |
| GAM | 0.865 | 0.892 | 0.854 |
| GBM | 0.863 | 0.891 | 0.851 |
| Lasso | 0.858 | 0.885 | 0.845 |
| SVM | 0.853 | 0.843 | 0.803 |
| RF | 0.849 | 0.875 | 0.844 |

Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.
Notes: The area under the ROC curve measures the overall predictive performance of the model. A machine learning algorithm that makes no predictive mistakes has an AUC of 1.0, while a model that predicts at random should achieve an AUC near 0.5.

**TABLE 9. True Positive Rate, False Positive Rate, and Accuracy of Models Predicting School Dropout**

| | *Scenario 1 10% of adolescents classified as dropouts* | | | *Scenario 2 30% of adolescents classified as dropouts* | | |
|---|---|---|---|---|---|---|
| *Dropout measure and MLA* | *True positive rate (sensitivity)* | *False positive rate (1 – specificity)* | *Accuracy* | *True positive rate (sensitivity)* | *False positive rate (1 – specificity)* | *Accuracy* |
| *A. Dropout in t + 1 or t + 2* | | | | | | |
| glmnet | 0.477 | 0.053 | 0.895 | 0.803 | 0.238 | 0.767 |
| GAM | 0.474 | 0.054 | 0.894 | 0.810 | 0.237 | 0.768 |
| GBM | 0.471 | 0.054 | 0.894 | 0.794 | 0.239 | 0.765 |
| Lasso | 0.461 | 0.055 | 0.891 | 0.789 | 0.239 | 0.764 |
| SVM | 0.449 | 0.057 | 0.889 | 0.801 | 0.238 | 0.766 |
| RF | 0.438 | 0.057 | 0.887 | 0.770 | 0.236 | 0.765 |
| *B. Dropout in t + 1* | | | | | | |
| glmnet | 0.567 | 0.067 | 0.909 | 0.879 | 0.259 | 0.750 |
| GAM | 0.584 | 0.066 | 0.911 | 0.874 | 0.260 | 0.749 |
| GBM | 0.561 | 0.068 | 0.908 | 0.881 | 0.259 | 0.750 |
| Lasso | 0.561 | 0.068 | 0.908 | 0.861 | 0.261 | 0.747 |
| SVM | 0.494 | 0.072 | 0.899 | 0.807 | 0.264 | 0.740 |
| RF | 0.539 | 0.069 | 0.905 | 0.833 | 0.252 | 0.754 |
| *C. Dropout in t + 2* | | | | | | |
| glmnet | 0.483 | 0.064 | 0.897 | 0.800 | 0.252 | 0.752 |
| GAM | 0.486 | 0.063 | 0.898 | 0.800 | 0.252 | 0.752 |
| GBM | 0.481 | 0.064 | 0.897 | 0.799 | 0.253 | 0.752 |
| Lasso | 0.476 | 0.064 | 0.896 | 0.794 | 0.253 | 0.751 |
| SVM | 0.434 | 0.068 | 0.889 | 0.718 | 0.260 | 0.738 |
| RF | 0.491 | 0.063 | 0.899 | 0.773 | 0.250 | 0.752 |

Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

Notes: I link each student in primary and secondary education who concluded their academic year $t$ and did not graduate from their secondary studies with him- or herself in years $t + 1$ and/or $t + 2$. Using this procedure, I identify the students who dropped out of school after year $t$ (in $t + 1$, $t + 2$, or either $t + 1$ or $t + 2$). Dropping out of school means failing to enroll or enrolling but withdrawing before the end of the academic year.

I derive the confidence intervals of the AUC for some of these algorithms (for dropout_t12). The difference between the glmnet and GBM models is not statistically significant at the 95 percent level. The AUC of both models is statistically significantly different from lasso.

Table 9 shows how the performance of these models translates into targeting effectiveness, based on their true positive rate, false positive rate, and accuracy. To set a common threshold for comparing the MLAs on these three measures, I define two scenarios: one in which 10 percent of adolescents are classified as future dropouts (those with the highest probability of dropping out in the future) and one in which 30 percent of adolescents are classified as future dropouts.[15] Under the first scenario, the glmnet algorithm has the best performance for the broadest measure of dropout (in either in $t + 1$ or $t + 2$), finding future dropouts at a rate of 477 out of 1,000. Additionally, it has a false positive rate of 0.053. In other words, nondropouts are incorrectly classified as dropouts at a rate of 53 cases out of 1,000. Finally, this algorithm successfully classifies 89.5 percent of the students. The second- and third-best-performing models in the first scenario are GAM and GBM, consistent with the AUC ranking. The true positive rates in these cases are 0.474 and 0.471, respectively. The false positive rate and the accuracy indicators are the same for both algorithms, at 0.054 and 0.894, respectively. The two algorithms with the lowest targeting performance are RF and SVM. The first of these algorithms finds future dropouts at a rate of 438 out of 1,000 and misclassifies nondropouts at a rate of 57 cases out of 1,000. These results are consistent with the ROC curves.

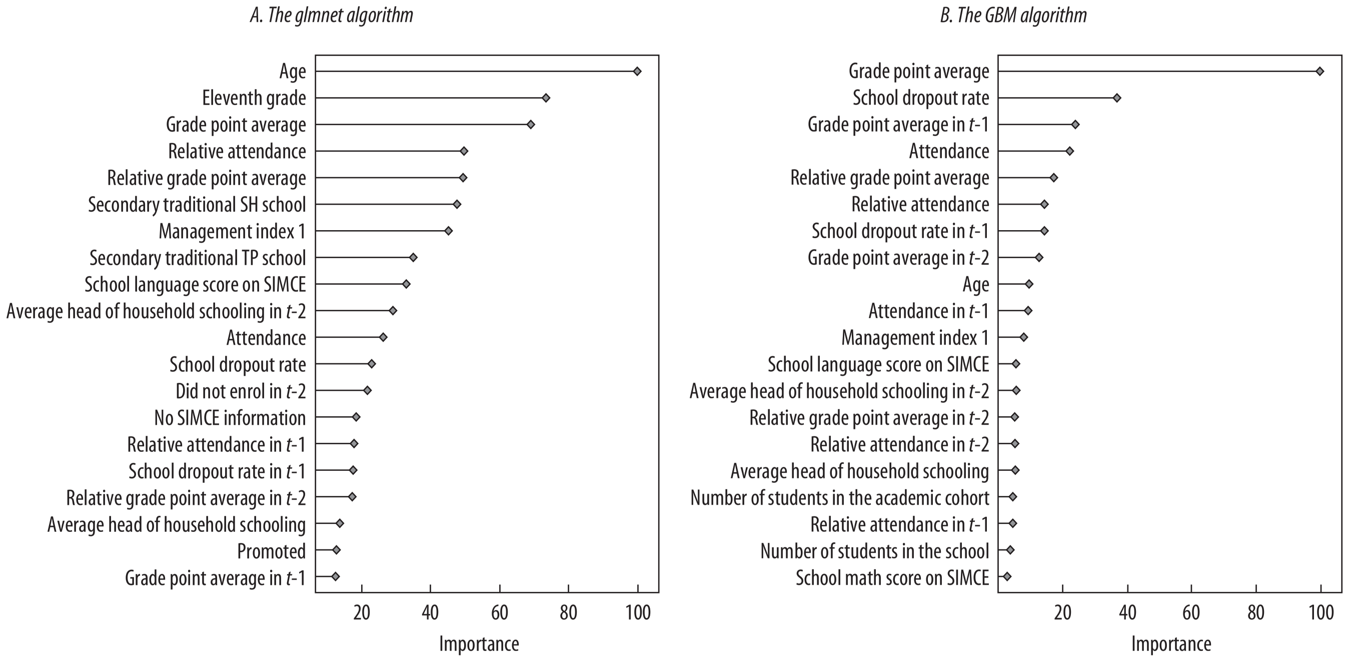
Under the second scenario, in which 30 percent of adolescents are classified as dropouts, the best performance belongs to the GAM algorithm. In this context, dropouts are found at a rate of 810 out of 1,000. However, the false positive rate and accuracy weaken, at 0.237 for the former and 0.768 for the latter. The algorithm based on elastic nets (glmnet) has the second-best performance after GAM on the true positive rate and accuracy. The results are similar for the narrower dropout measures, with some minor differences in ranking.

Figure 4 shows the most important variables for predicting school dropout for glmnet and GBM. The five most important differ for the two models and include age, grade point average in years $t$ and $t - 1$, attendance in year $t$, relative grade point average and attendance in year $t$, the student's grade level (seventh to twelfth) in year $t$, and the previous average dropout rate in the school. Per capita income plays a minor role in helping school dropout prediction in these two models.

Overall, the differences in performance across the models are small in magnitude. In general, glmnet, GAM, and GBM are the top performers, while SVM gives the worst results. The best MLAs produce adequate predictions of school dropout. Regarding the true and false positive rates, my results are better than or in the same region as 107 of the 110 dropout flags analyzed by Bowers, Sprott, and Taff (2013). The results provided by glmnet are better than those obtained for Guatemala and Honduras (Adelman and others, 2018). The accuracy levels under the first scenario in table 9, around 90 percent, are

15. Crespo (2019, appendix) shows the results of table 8 and table 9 using the second test data set, which assesses the predictions over time (out-of-time validation).

**F I G U R E 4 . Relative Importance of Variables**




Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

Notes: Variables are for year *t* unless otherwise noted. In the caret package, the maximum value of variable importance is 100. The procedure to estimate the variable importance varies by approach. For example, algorithms based on trees require permuting predictors to assess their accuracy while linear models utilize the absolute value of the *t* statistic of each coefficient in the regression.

equivalent to the results obtained by the best-performing MLA tested in North Carolina (Sorensen, 2019). My AUC findings are also in line with the best-performing models of school dropout tested in Wisconsin (Knowles, 2015), in which most of the algorithms had an AUC of between 0.860 and 0.870. However, these results are below the AUCs of 0.948 and 0.965 observed in Denmark (Şara, Halland, and Alstrup, 2015).

## Results: Targeting Assessment

This section presents the results of the targeting assessment. The first part provides summary statistics describing the relationship between poverty and school dropout and between each targeting mechanism and the last two outcomes. I then present the results for total leakage and leaked welfare.

### *Summary Statistics*

Table 10 provides bivariate summary statistics between targeting mechanisms (organized in quintiles) and the outcomes of the targeting assessment, namely, poverty status and school dropout. I offer both the mean value and the relative frequency. As poverty status and school dropout are dichotomous variables (equal to zero or one), the mean value can be interpreted as the proportion of poor adolescents and school dropouts in each quintile. The relative frequency describes the distribution of poor adolescents and future school dropouts among the quintiles. The targeting mechanisms included in the table are PMT scores in the SPF and predictions of the best-performing algorithm in the previous section (glmnet). The measure of school dropout I present is dropout_t12.[16]

As the table shows, there is a negative correlation between household per capita income and dropping out. The proportion of adolescents who leave school at any time within two years declines steadily from the first income quintile to the fifth, from fifteen out of 100 adolescents to only seven. Regarding the relative distribution of future school dropouts among the per capita income quintiles, 27.97 percent of adolescents who dropped out belonged to the first income quintile in the sample, versus 12.13 percent in the highest quintile. Accordingly, there is not a big overlap between poor adolescents and future school dropouts in my sample.

16. Summary statistics for the other two measures of school dropout are available in Crespo (2019, appendix).

**TABLE 10. Mean and Relative Frequency of Poor and School Dropout, by Quintile**

| | *Poor* | | *Dropout in t + 1 or t + 2* | |
|---|---|---|---|---|
| *Quintile* | *Mean* | *Relative frequency (%)* | *Mean* | *Relative frequency (%)* |
| *A. Per capita income* | | | | |
| First (lowest) | 1.00 | 100.00 | 0.15 | 27.97 |
| Second | 0.00 | 0.00 | 0.13 | 23.05 |
| Third | 0.00 | 0.00 | 0.11 | 19.81 |
| Fourth | 0.00 | 0.00 | 0.09 | 17.04 |
| Fifth (highest) | 0.00 | 0.00 | 0.07 | 12.13 |
| Total | 0.20 | 100.00 | 0.11 | 100.00 |
| *B. SPF scores* | | | | |
| First (lowest) | 0.44 | 44.03 | 0.14 | 25.59 |
| Second | 0.32 | 32.28 | 0.13 | 23.38 |
| Third | 0.16 | 16.29 | 0.12 | 20.94 |
| Fourth | 0.06 | 5.93 | 0.10 | 17.38 |
| Fifth (highest) | 0.01 | 1.47 | 0.07 | 12.71 |
| Total | 0.20 | 100.00 | 0.11 | 100.00 |
| *C. Predictive model of school dropout* | | | | |
| First (lowest) | 0.29 | 28.81 | 0.38 | 69.63 |
| Second | 0.24 | 24.06 | 0.10 | 18.76 |
| Third | 0.20 | 20.18 | 0.04 | 7.46 |
| Fourth | 0.16 | 16.39 | 0.02 | 3.11 |
| Fifth (highest) | 0.11 | 10.56 | 0.01 | 1.05 |
| Total | 0.20 | 100.00 | 0.11 | 100.00 |

Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

Notes: I define poverty as being in the first quintile of household per capita income in the sample. For SPF scores, the first quintile corresponds to adolescents in the bottom 20 percent of scores. The predictive model is the best-performing algorithm (glmnet), where the first quintile corresponds to adolescents with the 20 percent highest probability of dropping out of school.

In light of these results, it is likely that a targeting instrument designed to find one specific group (such as the poor) will have a lower capacity to identify the other group (school dropouts). The table shows a negative correlation between SPF scores and leaving school. To illustrate, fourteen out of 100 students in the bottom 20 percent of SPF scores dropped out, but only seven out of 100 did so among the fifth quintile of SPF scores. Also, there is an inverse relationship between PMT scores and poverty. For example, forty-four out of 100 adolescents in the first quintile of SPF scores are poor (defined as being in the first household income quintile), versus only one out of 100 students in the highest SPF quintile. Regarding relative frequencies, 25.59 percent of adolescents who dropped out belong to the first SPF quintile, and only 12.71 percent of dropouts are from the fifth SPF quintile.

The SPF score is a better tool for finding poor adolescents than for finding future dropouts. The first quintile of PMT scores encompasses 44.03 percent of poor adolescents, but only 25.59 percent of future dropouts. These findings are explained not by problems in the SPF model but rather by the low overlap between poverty and school dropouts in my sample.[17] In fact, the relative frequencies of school dropout by quintile are similar in magnitude for household income and SPF scores.

In contrast, the predictive model is more effective at finding future dropouts than poor adolescents: the first quintile captures 69.63 percent of school dropouts but only 28.81 percent of the poor. Regarding absolute values, there are more future dropouts than poor students in the first quintile of the predictive model. The latter is the case despite the population of future dropouts being smaller relative to poor adolescents.

I extract two key findings from the table. First, the SPF score is better than machine learning outputs at finding poor adolescents. In other words, using the PMT is more progressive than using the predictive model of school dropout. In the first quintile of the SPF, forty-four out of 100 students are poor, while in the first quintile of the predictive model, only twenty-nine out of 100 students are poor. Second, the PMT is less effective than the algorithm at finding future dropouts: the first quintile of SPF scores captures fourteen out of 100 dropouts; the first quintile of the predictive model, thirty-eight out of 100. Thus, prioritizing the use of SPF scores to target a CCT increases the effectiveness in terms of finding the poor but decreases the capacity to find future dropouts.

### *Targeting Assessment: Total Leakage*

This subsection presents the central results of the targeting assessment. For simplicity, the evaluation focuses on one measure of school dropout, namely, the indicator that captures whether an adolescent dropped out in year $t + 1$ or year $t + 2$ (dropout_t12). Thus, I use the outputs of the best MLA predicting dropout_t12 as a targeting mechanism (glmnet). The results for the other two measures of school dropout are available in Crespo (2019, appendix).

17. In 2013, only 33.2 percent of Chilean school dropouts aged fifteen to nineteen years (48.2 percent for men and 15.0 percent for women) dropped out for economic reasons (Salas Opazo, Ormazabal, and Crespo, 2015). Additionally, a survey conducted in ten Latin American cities among youth aged fifteen to twenty-five shows similar trends, with 44 percent of men and only 25 percent of women citing economic reasons for dropping out of school (Berniell and others, 2016). Other reasons for dropping out include lack of interest and adolescent parenthood, which might explain why there is not a high degree of overlap between poverty and school dropout. As countries progress economically, it seems likely that economic status will become less prevalent in explaining dropping out of school.

**TABLE 11. Targeting Indicators by Independent Approach and Available Budget**

| *SPF score only* | | *Predictive model only* | |
|---|---|---|---|
| *A. The budget allows a CCT to reach 5% of adolescents* | | | |
| Poor undercoverage | 0.867 | Poor undercoverage | 0.922 |
| Nonpoor leakage | 0.470 | Nonpoor leakage | 0.689 |
| Dropout undercoverage | 0.934 | Dropout undercoverage | 0.696 |
| Nondropout leakage | 0.855 | Nondropout leakage | 0.329 |
| Total leakage | 0.412 | Total leakage | 0.232 |
| *B. The budget allows a CCT to reach 20% of adolescents* | | | |
| Poor undercoverage | 0.560 | Poor undercoverage | 0.712 |
| Nonpoor leakage | 0.560 | Nonpoor leakage | 0.712 |
| Dropout undercoverage | 0.744 | Dropout undercoverage | 0.304 |
| Nondropout leakage | 0.859 | Nondropout leakage | 0.615 |
| Total leakage | 0.493 | Total leakage | 0.444 |
| *C. The budget allows a CCT to reach 40% of adolescents* | | | |
| Poor undercoverage | 0.237 | Poor undercoverage | 0.471 |
| Nonpoor leakage | 0.618 | Nonpoor leakage | 0.736 |
| Dropout undercoverage | 0.510 | Dropout undercoverage | 0.116 |
| Nondropout leakage | 0.865 | Nondropout leakage | 0.756 |
| Total leakage | 0.544 | Total leakage | 0.563 |

Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

Table 11 shows the results for the first two (out of four) targeting mechanisms, one based solely on the proxy means test score of the social protection file and one based exclusively on the outputs of the MLA. As the table shows, there is a trade-off between finding the poor and finding future dropouts. Under any budget scenario, poor undercoverage and nonpoor leakage increase when switching from SPF scores to the predictive model. For example, when the budget allows for providing the CCT to 5 percent of adolescents in the sample, poor undercoverage is 0.867 and nonpoor leakage is 0.470 if I use only the SPF for targeting. If I use the predictive model, these indicators increase to 0.922 and 0.689, respectively. Conversely, undercoverage of dropouts and leakage of nondropouts decrease when the output of glmnet replaces the PMT. To illustrate, when the budget allows for providing the CCT to 20 percent of students in the sample, dropout undercoverage is 0.744 and nondropout leakage is 0.859 when targeting is based on the SPF. If I use the predictive model for targeting, these indicators drop to 0.304 and 0.615, respectively.

An additional trade-off is related to expenditure. As the budget increases, undercoverage drops among both target groups, but leakage rates increase. Another finding is that the optimal targeting mechanism depends on the available budget. In the first two budget scenarios, total leakage is higher if the SPF

**TABLE 12. Targeting Indicators by Combined Approach and Available Budget**

| *25% SPF score, 75% model* | | *75% SPF score, 25% model* | |
|---|---|---|---|
| *A. The budget allows a CCT to reach 5% of adolescents* | | | |
| Poor undercoverage | 0.882 | Poor undercoverage | 0.852 |
| Nonpoor leakage | 0.530 | Nonpoor leakage | 0.408 |
| Dropout undercoverage | 0.740 | Dropout undercoverage | 0.854 |
| Nondropout leakage | 0.426 | Nondropout leakage | 0.679 |
| Total leakage | 0.158 | Total leakage | 0.232 |
| *B. The budget allows a CCT to reach 20% of adolescents* | | | |
| Poor undercoverage | 0.665 | Poor undercoverage | 0.609 |
| Nonpoor leakage | 0.665 | Nonpoor leakage | 0.609 |
| Dropout undercoverage | 0.349 | Dropout undercoverage | 0.524 |
| Nondropout leakage | 0.640 | Nondropout leakage | 0.737 |
| Total leakage | 0.419 | Total leakage | 0.442 |
| *C. The budget allows a CCT to reach 40% of adolescents* | | | |
| Poor undercoverage | 0.441 | Poor undercoverage | 0.298 |
| Nonpoor leakage | 0.720 | Nonpoor leakage | 0.649 |
| Dropout undercoverage | 0.138 | Dropout undercoverage | 0.243 |
| Nondropout leakage | 0.762 | Nondropout leakage | 0.791 |
| Total leakage | 0.553 | Total leakage | 0.508 |

Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

score is used (relative to the predictive model). However, when the budget allows the CCT to reach 40 percent of adolescents in the sample, total leakage is higher if the predictive model is used.

Table 12 shows the results for two combined targeting mechanisms. In the first, I allocate 25 percent of the budget based on the PMT score and the remaining 75 percent based on the MLA. In the second, the percentages are reversed, with the first 75 percent of the budget allocated via the SPF and the remaining 25 percent via the algorithm. The findings are similar to the single targeting mechanisms. First, there is a trade-off associated with the selection of the mechanism. Assigning a higher fraction of the budget based on the SPF translates into lower undercoverage of the poor and nonpoor leakage but a greater undercoverage of future dropouts and nondropout leakage. Second, when the budget increases, so do all the leakage rates, yet undercoverage decreases for both target groups. Third, the mechanism with the lowest total leakage depends on the budget at disposal.

Table 13 summarizes the total leakage indicator for the two independent mechanisms and the two combined mechanisms (from tables 11 and 12). For a fixed budget, table 13 identifies the targeting mechanism with the lowest total leakage. A combined approach is more effective at finding the poor or future

**TABLE 13. Total Leakage by Targeting Mechanism and Available Budget**

| | CCT coverage | | |
|---|---|---|---|
| *Targeting mechanism* | *x = 5%* | *x = 20%* | *x = 40%* |
| 0% SPF, 100% model | 0.232 | 0.444 | 0.563 |
| 25% SPF, 75% model | 0.158 | 0.419 | 0.553 |
| 75% SPF, 25% model | 0.232 | 0.442 | 0.508 |
| 100% SPF, 0% model | 0.412 | 0.493 | 0.544 |

Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

dropouts relative to an independent approach. For example, in the context where the budget allows reaching 5 percent of the sample, the mechanism that uses 25 percent of the SPF and 75 percent of the predictive model provides the lowest level of total leakage. In this example, only 15.8 percent of students who are assigned the hypothetical CCT are neither poor nor dropouts. In the other two budget scenarios, a mechanism that uses both sources of information also provides the optimal solution. In the second case, each combined mechanism performs better in the simulations than do the independent mechanisms. When the budget increases to 40 percent, the optimal mechanism within the alternatives I analyze is to allocate the first 75 percent of the resources using the PMT score.

I use multiple alternative specifications to test the robustness of these results. First, I change the definitions of the poverty line and income. Second, I modify the measure of school dropout. Third, I use an alternative combined approach, consisting of a single composite score derived from weighting both instruments (the SPF and the predictive model) and assigning the hypothetical CCT using this new index. Finally, I replace the best MLA (glmnet) with boosted trees (GBM) and lasso. Overall, my findings are robust to these alternative specifications. A targeting mechanism that uses the PMT score in conjunction with the predictive model minimizes total leakage (relative to independent mechanisms) in every scenario. This finding does not change depending on the budget, poverty line, income definition, dropout measure, or algorithm used.[18]

In practice, changing the targeting mechanism of a CCT from a PMT alone to a mechanism that also requires using a predictive model of school dropout implies new targeting costs. However, even after I add administrative costs to the specifications, a targeting approach that relies on both sources

18. The robustness results are available in Crespo (2019, appendix).

**TABLE 14. Leaked Welfare by Social Valuation of Target Groups**

| *Social priority and targeting mechanism* | *CCT coverage* | | | | | |
|---|---|---|---|---|---|---|
| | *Twice as important* | | | *Four times more important* | | |
| | *x = 5%* | *x = 20%* | *x = 40%* | *x = 5%* | *x = 20%* | *x = 40%* |
| *A. The poor are more important than dropouts* | | | | | | |
| 0% SPF, 100% model | 0.569 | 0.680 | 0.742 | 0.617 | 0.692 | 0.740 |
| 25% SPF, 75% model | 0.495 | 0.657 | 0.734 | 0.509 | 0.660 | 0.729 |
| 75% SPF, 25% model | 0.498 | 0.652 | 0.696 | 0.462 | 0.635 | 0.677 |
| 100% SPF, 0% model | 0.599 | 0.659 | 0.700 | 0.547 | 0.619 | 0.668 |
| *B. Dropouts are more important than the poor* | | | | | | |
| 0% SPF, 100% model | 0.449 | 0.647 | 0.749 | 0.401 | 0.634 | 0.752 |
| 25% SPF, 75% model | 0.461 | 0.649 | 0.748 | 0.447 | 0.645 | 0.753 |
| 75% SPF, 25% model | 0.589 | 0.694 | 0.743 | 0.625 | 0.711 | 0.762 |
| 100% SPF, 0% model | 0.727 | 0.759 | 0.783 | 0.778 | 0.799 | 0.815 |

Source: Author's calculations, using administrative data sets from the Chilean Ministry of Education and Ministry of Social Development.

of information remains more effective than an independent approach. This holds for all combinations of fixed and variables costs added to targeting mechanisms that incorporate the predictive model.[19]

## *Targeting Assessment: Leaked Welfare*

Unlike total leakage, leaked welfare is affected by differences in the social valuation of target groups. For example, if the poor are twice as important as future dropouts, then leaked welfare is zero if all hypothetical recipients of a CCT are poor and future dropouts, one if all beneficiaries are nonpoor and not future dropouts, one-third if all potential recipients are poor but not future dropouts, and two-thirds if all potential recipients are future dropouts but nonpoor. If finding future school dropouts is four times more important than finding the poor, leaked welfare is zero if all recipients are future school dropouts and poor, one if they are neither poor nor future dropouts, one-fifth if all of them are future school dropouts but are not poor, and four-fifths if they are all poor but not future school dropouts.

Table 14 presents the results of the assessment when the social valuation of the target groups differs. Panel A shows that when the poor are valued more highly than future dropouts, it is beneficial to make extensive use of the social protection file to select beneficiaries. The combined mechanism that assigns the first 75 percent of the budget using the SPF provides the lowest leaked

19. See Crespo (2019, appendix).

welfare in two out of three scenarios when the poor are twice as important. When the poor are four times more important, the optimal mechanism in two out of three scenarios is to use the PMT score exclusively.

Panel B demonstrates that relying exclusively on the predictive model is mostly the optimal mechanism when future dropouts are valued more highly than the poor. When dropouts are four times more important than the poor, not using the SPF minimizes leaked welfare in all three budget scenarios. A combined mechanism is superior only when a large budget is available and targeting dropouts is twice as important as the poor.

Overall, the leaked welfare measure I provide in this subsection improves our understanding of the targeting performance of different mechanisms. When the social valuation of the target groups differs to a large extent, the preferred mechanism is the one designed to find the target group that is most socially valued. When the welfare weight, $\gamma$, assigned to a future dropout is much higher than that of a poor student, using solely the predictive model is the optimal mechanism to maximize welfare. Conversely, when finding a poor adolescent has a much higher $\gamma$ than finding a future dropout, prioritizing the PMT mostly provides higher levels of welfare.

## *Concluding Remarks*

The development of quality administrative records has expanded the possibilities for improving program design and conducting cost-effective research. Within a big data context, this paper contributes a general methodology to improve targeting design and assessment when two or more target groups matter and there are trade-offs between potential targeting mechanisms.

This paper offers targeting indicators that combine information on a program's key target groups. In this context, the paper has analyzed whether a proxy means test and alternative targeting mechanisms based on a predictive model of school dropout, built with machine learning algorithms, are effective tools for reaching both the poor and future school dropouts. Overall, the paper provides novel contributions to the policy-targeting field. The paper's findings transcend the specific Chilean CCT case used to analyze the targeting mechanisms. More generally, the findings are relevant for countries that either wish to develop predictive models using administrative records or want to strengthen the targeting of their policies where multiple target groups exist.

The results of the CCT assessment show that a trade-off exists between using the PMT versus the predictive model. Using the PMT for targeting is more progressive, as poor undercoverage and nonpoor leakage are reduced, but future dropout undercoverage and nondropout leakage increase. This trade-off is explained by the low level of overlap between poverty and school dropout in Chile. Generally, it is more effective to use these two mechanisms in conjunction than to use them independently. However, another key finding is that the use of a combined approach is not necessarily more effective when the social valuation of the two target groups differs to a large extent. Thus, the combined targeting approach is likely to be useful for policymakers in countries where multiple groups are relevant for targeting and where there is far from perfect overlap among the groups. This is also likely to hold for higher dimensions if finding at least one target group is considered successful targeting, though further research is needed to analyze this in detail.

My results are partly in line with the findings of Azevedo and Robles (2013). My paper and theirs both offer a multidimensional targeting approach that fosters the notion that more than one target group and more than one targeting criterion should exist for CCT design and assessment. However, in their case no trade-offs exist between the targeting mechanisms assessed and the relevant target groups. By offering two metrics that process these trade-offs, my paper facilitates decision-making in targeting design.

Regarding CCT policy implications, my paper advances the idea that CCT targeting can be improved when other dimensions beyond income are considered. This finding invites policymakers to broaden the targeting design by adding the human capital accumulation dimension. Achieving a better balance among target groups in CCT allocation could also help to enrich and diversify the targeting assessment of these schemes, where a unidimensional outlook has prevailed (Maluccio, 2009; Robles, Rubio, and Stampini, 2015; Skoufias, Davis, and de la Vega, 2001; Stampini and Tornarolli, 2012). An essential and implicit takeaway from the paper is that effective targeting depends on consistency. Targeting design must reflect the goals of the policy and the consequential definition of the target groups. If a cash transfer has multiple target groups, then unidimensional targeting may not be the most effective design for the program.

The latter conclusion does not necessarily hold if public officials strongly prioritize finding the poor over other target groups. In this case, maintaining the status quo—namely, targeting CCTs based on income—is appropriate. Alternatively, policy designers should evaluate the cost-effectiveness of adopting a new targeting mechanism for CCTs. A first step in this sense would be to estimate the costs of developing and implementing a new targeting mechanism, estimate the gains in targeting effectiveness, and then compare these with the default scenario.

Using the framework of social welfare models can enrich the discussion of effective targeting for CCTs. In theory, CCTs should prioritize the groups where the impact is largest. These are the poorest among the poor and adolescents who would drop out of school in the absence of the CCT. For example, a CCT might have little impact on an adolescent with little motivation to continue studying because of low school quality. In practice, though, targeting CCTs using these criteria requires not only a flawless measurement of the degrees of poverty but also a perfect understanding of the causes of potential school dropout for each adolescent.

Building on this paper, future research could strengthen my social welfare analysis. One limitation of my targeting assessment is that I use only undercoverage and leakage rates. For example, I make no distinction between those at the bottom of the distribution and those who are marginally poor. I have assumed that the social valuation of finding any poor is the same. A similar shortcoming exists in the case of dropouts. This analysis could also be enriched if the size of the transfer varies, since higher transfers increase the likelihood of obtaining the desired effects. Additional angles for future research along these lines are to include more dimensions than education (such as health, by including children who are not attending preventive check-ups as a target group); to consider other stages in the educational cycle (such as preschool); and to use new models or means tests instead of the PMT used in this paper.

Another distinctive contribution of my paper is the predictive model of school dropout. The literature is extensive on the topic of determinants, but less so on predictions. The core of this research comes from developed countries, especially the United States. My paper is one of the first, along with that of Adelman and others (2018), to use large administrative data sets outside a developing nation to study this topic. Furthermore, there are not many applications of MLAs for school dropout. The most effective algorithms produce results that are in line with the related literature (Adelman and others, 2018; Knowles, 2015; Sorensen, 2019) and that are better than most of the dropout flags analyzed by Bowers, Sprott, and Taff (2013). The best model in predicting school dropout at any point within two years reaches an area under the ROC curve of 0.866.

These results have important policy implications as they show that appropriate predictive models of school dropout using administrative data sets are at hand for public officials. Naturally, the selection of variables is restricted by the availability of administrative records, given that the models I implement rely solely on information that is currently available from the Chilean government. No variables are provided by costly surveys. This finding has policy implications for every policy that defines students at risk of dropping out of school as their target group. For example, the impact of early warning systems could be improved by strengthening their ability to find students who are more likely to drop out of school. In contexts where countries are improving their administrative records, these lessons deserve attention.

Future research could also test longitudinal and multilevel models for the predictive approach. In fact, my approach to finding *f* does not precisely match that of Lamote and others (2013), which is in the category of longitudinal multilevel modeling. Longitudinal growth models have provided the most accurate predictions on school dropout (Bowers, Sprott, and Taff, 2013). Another potential direction for further research would be to improve the capabilities of the predictive model of school dropout by adding new variables. For example, in Chile, pregnancy and motherhood are relevant drivers of school dropout (Salas Opazo, Ormazabal, and Crespo, 2015). Young mothers can be identified through the administrative data from the Civil Registry Office and added to the predictive model. Additionally, the frequency of some predictors I use in my model could be enhanced. For example, the Chilean Ministry of Education has monthly attendance records at the individual level. This information could be useful if attendance levels in the last months of an academic year are a stronger predictor of future dropout than attendance when an academic year starts. A variable measuring absences in the last month is used by the top-performing algorithm in predicting high school dropout in the literature (Şara, Halland, and Alstrup, 2015).

CCTs continue to be a relevant social policy across the globe. Their goals of poverty alleviation and human capital accumulation remain valid in multiple countries. This paper aims to improve the design and assessment of their targeting. In Chile, a country with rich administrative data sets, using a PMT in conjunction with a predictive model of school dropout allows for finding more adolescents who are either poor or future school dropouts. Public officials who value these two target groups equally may find opportunities for increased targeting effectiveness by modifying the allocation rules of CCTs.

More generally, when multiple target groups or dimensions are relevant, policymakers and evaluators can benefit by adopting targeting designs and assessments along the lines discussed. Policymakers' decisions can be optimized using data-driven approaches. This paper provides a flexible framework for targeting, tailored to a big data context, when multiple groups matter.

## References


Abadie, Alberto Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2014. "Finite Population Causal Standard Errors." Working Paper 20325. Cambridge, Mass.: National Bureau of Economic Research.

Adelman, Melissa, Francisco Haimovich, Andres Ham, and Emmanuel Vazquez. 2018. "Predicting School Dropout with Administrative Data: New Evidence from Guatemala and Honduras." *Education Economics* 26 (4): 356–72.

Athey, Susan, and Guido W. Imbens. 2015a. "A Measure of Robustness to Misspecification." *American Economic Review* 105 (5): 476–80.

———. 2015b. "Machine Learning Methods for Estimating Heterogeneous Causal Effects." Research Paper 3350. Stanford University, Graduate School of Business.

Attanasio, Orazio, Emla Fitzsimons, Ana Gomez, and others. 2010. "Children's Schooling and Work in the Presence of a Conditional Cash Transfer Program in Rural Colombia." *Economic Development and Cultural Change* 58 (2):181–210.

Azevedo, Viviane, and Marcos Robles. 2013. "Multidimensional Targeting: Identifying Beneficiaries of Conditional Cash Transfer Programs." *Social Indicators Research* 112 (2): 447–75.

Barrera-Osorio, Felipe, Marianne Bertrand, Leigh L. Linden, and Francisco Pérez-Calle. 2011. "Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia." *American Economic Journal: Applied Economics* 3 (2): 167–95.

Bassi, Marina, Matias Busso, and Juan Sebastián Muñoz. 2015. "Enrollment, Graduation and Dropout Rates in Latin America: Is the Glass Half Empty or Half Full?" *Economía* 16 (1): 113–56.

Baulch, Bob. 2002. "Poverty Monitoring and Targeting Using ROC Curves: Examples from Vietnam." Working Paper 161. University of Sussex, Institute of Development Studies.

Berniell, Lucila, Dolores de la Mata, Raquel Bernal, and others. 2016. "Mas habilidades para el trabajo y la vida: Los aportes de la familia, la escuela, el entorno y el mundo laboral." Reporte de Economía y Desarrollo. CAF Banco de Desarrollo de América Latina.

Bowers, Alex J., Ryan Sprott, and Sherry Taff. 2013. "Do We Know Who Will Drop Out? A Review of the Predictors of Dropping Out of High School: Precision, Sensitivity and Specificity." *High School Journal* 96 (2): 77–100.

Coady, David, Margaret Grosh, and John Hoddinott. 2004. *Targeting of Transfers in Developing Countries: Review of Lessons and Experience.* Washington, D.C.: World Bank.

Coady, David, and Emmanuel Skoufias. 2004. "On the Targeting and Redistributive Efficiencies of Alternative Targeting Instruments." *Review of Income and Wealth* 50 (1): 11–27.

Comité de Expertos de la Ficha de Protección Social. 2010. "Informe final: Comité de Expertos Ficha de Protección Social." Santiago, Chile: Ministerio de Planificación y Cooperación (MIDEPLAN).

Crespo, Cristian. 2019. "Two Become One: Improving the Targeting of Conditional Cash Transfers with a Predictive Model of School Dropout." Social Policy Working Paper 05-19. London: London School of Economics.

De Wachter, Stefan, and Sebastian Galiani. 2006. "Optimal Income Support Targeting." *International Tax and Public Finance* 13 (6): 661–84.

Fiszbein, Ariel, and Norbert Schady. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty.* Washington, D.C.: World Bank Group.

Galiani, Sebastian, and Patrick J. McEwan. 2013. "The Heterogeneous Impact of Conditional Cash Transfers." *Journal of Public Economics* 103:85–96.

Glewwe, Paul. 1992. "Targeting Assistance to the Poor: Efficient Allocation of Transfers When Household Income Is Not Observed." *Journal of Development Economics* 38 (2): 297–321.

Glewwe, Paul, and Pedro Olinto. 2004. "Evaluating the Impact of Conditional Cash Transfers on Schooling: An Experimental Analysis of Honduras' PRAF-II Program." Final Report for USAID. Washington, D.C.: International Food Policy Research Institute.

Grosh, Margaret E., and Judy L. Baker. 1995. "Proxy Means Tests for Targeting of Social Programs: Simulations and Speculation." Living Standards Measurement Study (LSMS) Working Paper 118. Washington, D.C.: World Bank Group.

Handa, Sudhanshu, and Benjamin Davis. 2006. "The Experience of Conditional Cash Transfers in Latin America and the Caribbean." *Development Policy Review* 24 (5): 513–36.

Honorati, Maddalena, Ugo Gentilini, and Ruslan G. Yemtsov. 2015. *The State of Social Safety Nets.* Washington, D.C.: World Bank Group.

Hunt, Frances. 2008. "Dropping Out from School: A Cross-Country Review of Literature." Consortium for Research on Educational Access, Transitions and Equity (CREATE). University of Sussex, Center for International Education.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95.

Knowles, Jared E. 2015. "Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin." *Journal of Educational Data Mining* 7 (3): 18–67.

Kuhn, Max. 2008. "Building Predictive Models in R Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26.

Lamote, Carl, Jan Van Damme, Wim Van den Noortgate, and others. 2013. "Dropout in Secondary Education: An Application of a Multilevel Discrete-Time Hazard Model Accounting for School Changes." *Quality and Quantity* 47 (5): 2425–46.

Maluccio, John A. 2009. "Household Targeting in Practice: The Nicaraguan Red de Protección Social." *Journal of International Development* 21 (1): 1–23.

Maluccio, John A., and Rafael Flores. 2005. "Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan Red de Protección Social." Research Report 141. Washington, D.C.: International Food Policy Research Institute.

McBride, Linden, and Austin Nichols. 2016. "Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning." *World Bank Economic Review* 32 (3): 531–50.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Nichols, Austin. 2018. "Implementing Machine Learning Methods in Stata." Paper presented at the 2018 London Stata Conference.

Peyre Dutrey, Alexander. 2007. "Successful Targeting? Reporting Efficiency and Costs in Targeted Poverty Alleviation Programmes." Social Policy and Development Programme Paper 35. Geneva: United Nations Research Institute for Social Development.

Ravallion, Martin, and Kalvin Chao. 1989. "Targeted Policies for Poverty Alleviation under Imperfect Information." *Journal of Policy Modelling* 11 (2): 213–24.

Robles, Marcos, Marcela G. Rubio, and Marco Stampini. 2015. "Have Cash Transfers Succeeded in Reaching the Poor in Latin America and the Caribbean." Policy Brief IDB-PB-246. Washington, D.C.: Inter-American Development Bank.

Rumberger, Russell W., and Sun Ah Lim. 2008. "Why Students Drop Out of School: A Review of 25 Years of Research." California Dropout Research Project 15. University of California, Santa Barbara, Gevirtz Graduate School of Education.

Salas Opazo, Victor, Claudia Ormazabal, and Cristián Crespo. 2015. "Informe Final de Evaluación: Programa Beca de Apoyo a la Retención Escolar." Santiago, Chile: Ministerio de Educación, Junta Nacional de Auxilio y Becas (JUNAEB).

Şara, Nicolae-Bogdan, Rasmus Halland, and Stephen Alstrup. 2015. "High-School Dropout Prediction Using Machine Learning: A Danish Large-Scale Study." In *Proceedings: 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, edited by Michel Verleysen, pp. 319–24. Louvain-la-Neuve, Belgium: Ciaco.

Schady, Norbert, and Maria Caridad Araujo. 2008. "Cash Transfers, Conditions, and School Enrollment in Ecuador." *Economía* 8 (2): 43–70.

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74 (1): 199–250.

Skoufias, Emmanuel, Benjamin Davis, and Sergio de la Vega. 2001. "Targeting the Poor in Mexico: An Evaluation of the Selection of Households into PROGRESA." *World Development* 29 (10): 1769–84.

Sorensen, Lucy C. 2019. "'Big Data' in Educational Administration: An Application for Predicting School Dropout Risk." *Education Administration Quarterly* 55 (3): 404–46.

Stampini, Marco, and Leopoldo Tornarolli. 2012. “The Growth of Conditional Cash Transfers in Latin America and the Caribbean: Did They Go Too Far?” Policy Brief IDB-PB-185. Washington, D.C.: Inter-American Development Bank.

Theodoridis, Sergios, and Konstantinos Koutroumbas. 2009. *Pattern Recognition.* Burlington, Mass.: Academic Press.

Universidad del Desarrollo. 2014. “Informe final: Evaluación de impacto de la bonificación ingreso ético familiar del Ministerio de Desarrollo Social (ex MIDEPLAN).” Santiago, Chile.

Van der Vaart, Aad W., Sandrine Dudoit, and Mark J. Van der Laan. 2006. “Oracle Inequalities for Multi-Fold Cross Validation.” *Statistics and Decisions* 24 (3): 351–71.

Varian, Hal R. 2014. “Big Data: New Tricks for Econometrics.” *Journal of Economic Perspectives* 28 (2): 3–28.

Wodon, Quentin T. 1997. “Targeting the Poor Using ROC Curves.” *World Development* 25 (12): 2083–92.