imds
instituto mobilidade e
desenvolvimento social

Institute for Mobility and Social Development

Dropout predictor for the final years of junior high school in the municipal network of Rio de Janeiro – Challenges and Lessons Learned.[i]

Rio de Janeiro, December 20, 2023

Introduction

Brazil, over the last decades, has faced persistent challenges related to school abandonment and dropout in elementary and junior high school. This problem, which directly affects the social and economic development of the country, has multifaceted roots, encompassing socioeconomic, pedagogical, and structural factors. Despite government efforts to universalize access to education, student retention in schools remains a critical obstacle, with impacts that extend beyond school walls, reflecting on the labor market and society as a whole.

School dropout is a complex issue that requires a multidimensional approach, incorporating the understanding of the variables that influence the student's decision to drop out of school. Among these factors, the quality of teaching, the family context, infrastructure issues and educational resources, as well as the engagement and perceived relevance of the school curriculum for the student's life stand out. A predictor of school dropout can be a differentiated tool in the early identification of risk signs, allowing targeted and effective interventions that can alter life trajectories and strengthen education as a pillar of the social mobility process.

The project Dropout Predictor for the final years of junior high school in the municipal network of Rio de Janeiro aims to act as a transforming agent, strategically inserting itself in this scenario. The need to intervene since elementary and junior high school is crucial, as the dropout event is often due to a cumulative process that begins in early childhood and ends up occurring at some point during basic education. Abandonment at this stage not only compromises access to education, but also impacts the integral formation of individuals, perpetuating inequalities and limiting future opportunities. With the purpose of becoming another front of information to support the actions of the *Bora Pra Escola* program, the project seeks to anticipate these adverse trajectories through the implementation of an alert system in the final years, fed by previous data from students that signal greater or lesser chances of dropping out before completing basic education.

The project developed throughout 2022 and 2023 within the scope of the TCA (Technical Communication Agreement) between the Institute for Mobility and

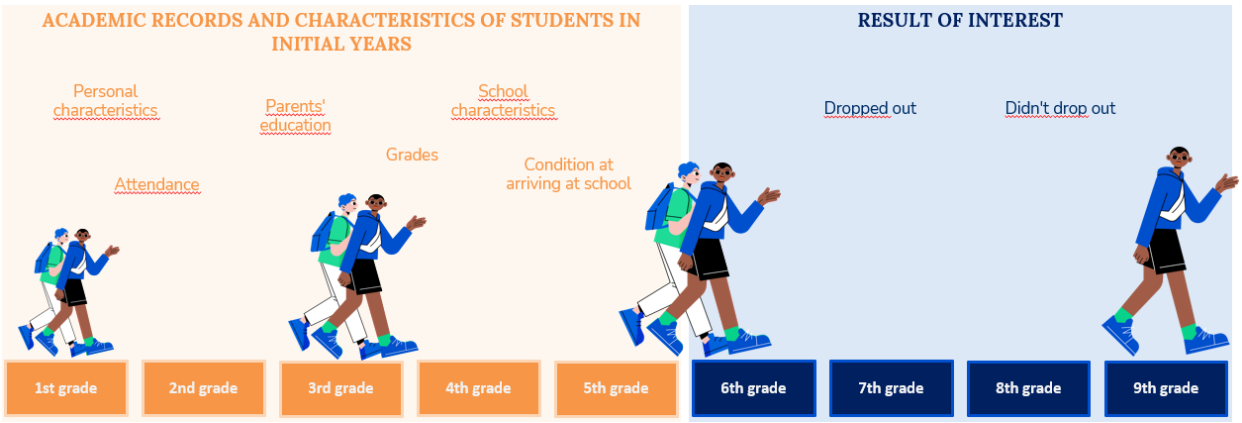
Social Development (IMDS) and the Municipal Department of Education of Rio de Janeiro had a series of challenges. The dropout predictor has achieved a performance below expectations and requires that new research stages be explored so that it actually becomes a differentiated tool in supporting the fight against dropout. This report presents all the stages of development of the predictor, the performance achieved, the challenges faced and especially the lessons learned so far, so that this information can be used for project continuity.

The Initial Project

The initial proposal consisted of creating a predictor of the probability of dropout, based on statistical modeling. The choice of the 5th grade as a point of strategic analysis is justified by its relevance in the educational path, being a decisive period that precedes the transition to the next stage. The objective of the project was to predict the chance of a student dropping out before completing elementary and junior high school.

As a concrete result, it was expected to obtain an annual list of students who completed the 5th grade, accompanied by the respective probability of dropping out by the end of junior high school.

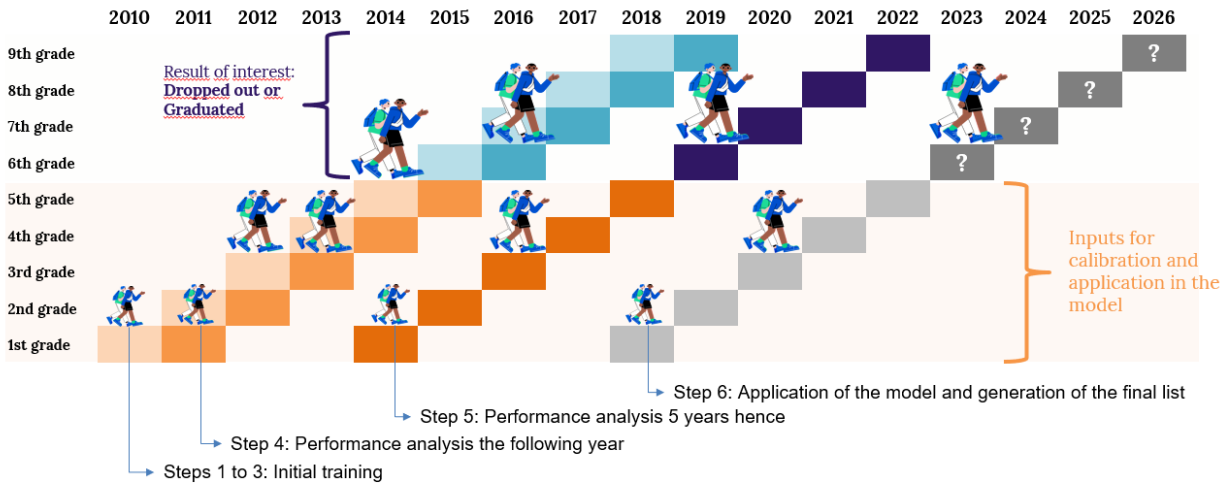
To achieve this objective, inputs such as the student's school trajectory, personal characteristics, parents' education, and academic records would be used, made available annually by SME/RJ. The following figure illustrates the initial proposal of the project.



The project adopted a six-step methodology that combines observation of

academic records, machine training and application of the model in successive cohorts:

- **Step 1:** Observation of the academic records and characteristics of students who have already passed through the network.
- **Step 2:** Relationship of the school records and characteristics of the students in the early years with the result of interest in the final years (dropped out or not).
- **Step 3:** Machine training to identify patterns between the information observed in the early years and the result of interest in the later years.
- **Step 4:** Verification of the performance of the trained model by applying it to the next cohort and evaluating the level of accuracy.
- **Step 5:** Application of the model in a more distant cohort to perform the final performance test.
- **Step 6:** Prediction (inference) applying the corrected parameters to generate a list of students with their respective dropout probabilities.



The predictor

Variables used

The central information of interest for the predictor is the probability of dropping out before the end of junior high school.

The variable chosen to represent this phenomenon was the record of abandonment as the last known movement information in all years, in agreement with the SME on the best possible way to identify it in the available administrative records. Thus, dropout was defined as 1 when the student had the last information on movement on the network as "Abandonment" and 0 for all other cases. This more comprehensive approach allows for a more accurate identification of dropout cases.

The following table shows the number of records analyzed and the last identified movement situation.

Situation	Last_move	Quantity
Inactive	Abandonment	577,284
Inactive	Completed level of education	456,227
Inactive	Enrollment renovation	365,669
Inactive	Transfer to the same network	353,090
Inactive	Transfer to private network	262,431
Inactive	Change of class	226,800
Inactive	Relocation to own network	214,765
Inactive	Transfer to state/federal public network	202,208
Inactive	Transfer to public network of other municipalities	160,961
Inactive	End of PEJA	144,863
Inactive	Needs to work	80,230
Inactive	Initial System Load	60,901
Inactive	Transfer within own network	42,292
Inactive	Duplicity in enrollment	23,720
Inactive	Initial enrollment	20,534
Inactive	Vacancy withdrawal	16,265
Inactive	Renewal	16,210
Inactive	Relocation from own network	10,803
Inactive	Transfer from private network	10,029
Inactive	Serious illness or abnormality	6,262
Inactive	Transfer from state/federal public network	5,529
Inactive	Adequacy of enrollment	5,081
Inactive	Death	4,805
Inactive	Regularization of COC resource	3,192
Inactive	Transfer from public network of other municipalities	2,984
Inactive	Change of teaching modality	2,495
Inactive	Reclassification (Entrance)	1,932
Inactive	Change of PEJA block	423
Inactive	Adequacy of enrollment	202
Inactive	Reclassification	1

The predictor variables were extracted from the class and assessment files, as described below:

About the student:

- Education of the 1st guardian
- Education of the 2nd guardian
- Regular or irregular trajectory
- Attendance in 5th grade
- Grades in 5tb grade

About the class:

- Last year in which they attended 5th grade.
- Number of changes of schools
- Number of failures

About the school:

- Management Complexity Index
- Percentage of students at each socioeconomic level (defined by INSE)

Data availability and quality

The process of accessing and criticizing data was a step that required a lot of effort and time, especially due to common problems in working with administrative records and the care necessary to comply with all LGPD[1] standards.

The SME had already been carrying out a long work of systematizing and standardizing the information over 10 years of students' history and organizing this data in a Data Warehouse so that it could be used in research studies to support the improvement of the programs and actions of the secretariat.

The data were made available and underwent a process of critical and feasibility analysis for the statistical modeling of interest. The first step was to verify the premise that the variables are approximately iid (independent and identically distributed) over time, when comparing the cohorts. When this does not occur, this phenomenon is called Covariate Shift. The procedure is composed of the selection of two subsets of data that are 4 years apart and the assignment of the target variable the value 1 to all records of the most recent year and 0 to those of the oldest year. Then, training a classification model, using the target variable created in the previous step.

To measure whether the trained model is able to identify the records belonging to each year using only the independent variables as a reference, the ROC-AUC (area under the receiver operator curve) metric was used, which varies between 0.5 and 1.0, where a value of 0.5 indicates that the years are indistinguishable (ideal case), using only the independent variables, and 1.0 indicates that the years are perfectly separable (worst case). By training the model with the data received by the SME/RJ and using all the independent variables, we obtained an ROC-AUC value above 0.9, indicating that the distribution of at least one independent variable in question changes over time.

[1] The acronym refers to Brazil's General Data Protection Law (LGPD).

This result implies the need for careful selection of the independent variables to be used in the model, in order to mitigate this effect and not obtain a low performance of the final model. After the process of selecting variables by re-executing the procedure described above for numerous subsets of independent variables, a subset was obtained that had the lowest possible degree of Covariate Shift for the set of data provided, taking into account the tradeoff of performance loss in the prediction of dropout. Consequently, a part of the variables was used in the model and another part was discarded because it did not have a good consistent predictive power over time to predict dropout. The following table summarizes the variables available and those that were effectively utilized.

Available and used variables	Available and unused variables
About the academic record:	About the student:
Last entry: if abandonment, indicates dropout	Sex
About the student:	Place of birth
Education of 1st guardian	Race/skin color
Education of 2nd guardian	Type of transport
Regular or irregular trajectory	Age in 5th grade
About assessments:	Nationality
Attendance in 5th grade	Impaired
Grades in 5th grade	About assessments:
About the class:	1st to 4th grade attendance
Last year attended = 5th grade	Grades from 1st to 4th grade
Number of school changes	About the class:
Number of failures	Number of mid-year school changes
About the school:	Number of class and shift changes
Management Complexity Index	About the school:
Percentage of students in each INSE	INSE

The analysis of the data identified the need to divide the file into two parts, considering students with "regular trajectories" in one group and students with "irregular trajectories" in another group. The irregular trajectories were defined by the identification of some atypical interruption in the process of annual advancement from one grade to another. The following table exemplifies the difference between the two types of trajectories:

Examples of "Regular trajectories"

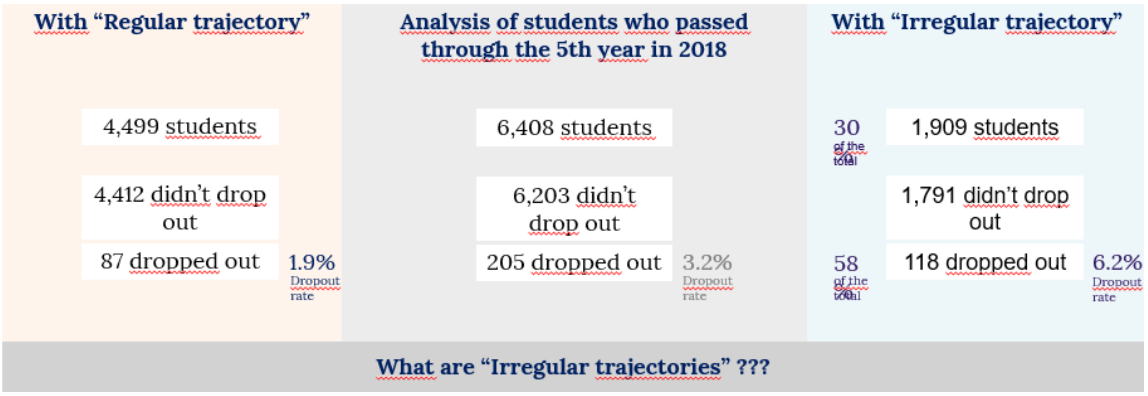
1st year
2nd year
3rd year
4th year
5th year
6th year
7th year
8th year
9th year

Examples of "Irregular trajectories"

year	student_id	grup
2000	X	Grup I
2001	X	Grup I
2001	X	Grup I
2002	X	1st Cycle of Education - Initial Period
2003	X	1st Cycle of Education - Intermediary Period
2003	X	1st Cycle of Education - Intermediary Period
2004	X	1st Cycle of Education - Final Period
2004	X	1st Cycle of Education - Final Period
2005	X	3rd Grade
2005	X	3rd Grade
2006	X	4th Grade
2006	X	4th Grade
2008	X	3rd Cycle of Education - Initial Period
2009	X	8th grade

year	student_id	grup
2007	Y	Nursery
2008	Y	Grup I
2009	Y	Grup I
2010	Y	1st grade
2011	Y	2nd grade
2012	Y	3rd grade
2013	Y	Re-literacy 1
2014	Y	4th grade
2015	Y	5th grade
2016	Y	6th grade
2019	Y	Carioca I
2020	Y	Carioca II

The division between the groups was necessary due to the identification that the group of students with regular trajectories had a much lower occurrence of dropout than the other group and, therefore, made the prediction exercise much more challenging. The table below shows the difference between the occurrence of dropout in the two groups considering a random sample of students.



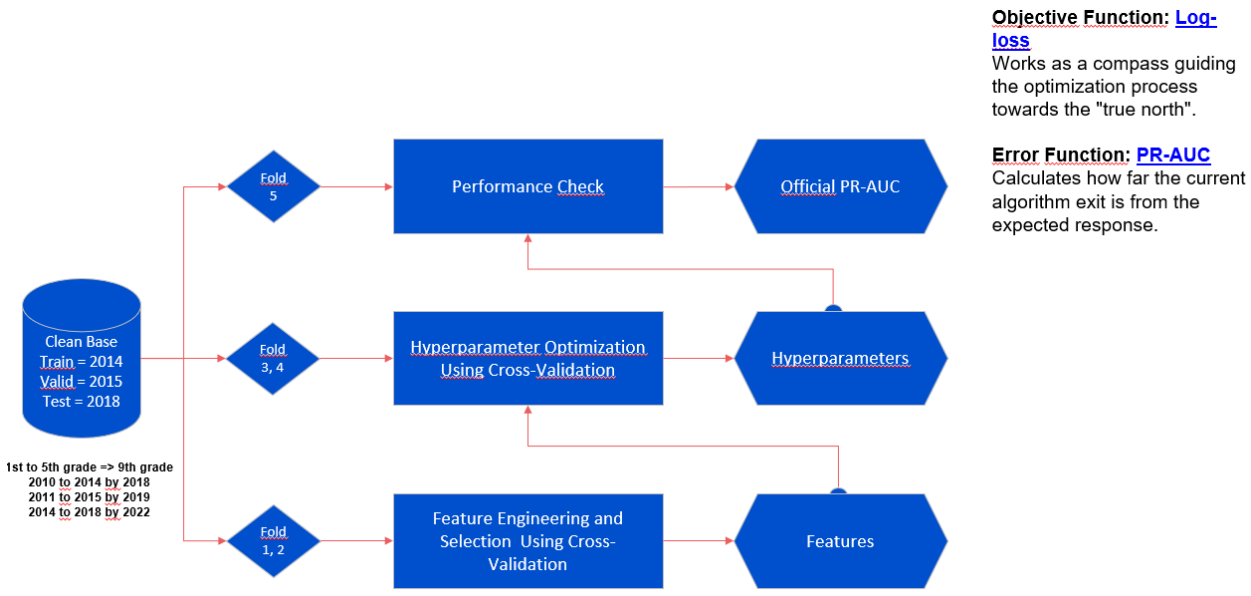
A second problem identified in the data analysis was the difference in the percentage of dropout identified in the administrative records and in the published data of INEP. This problem is possibly due to the lack of information about students when they are transferred to other networks before completing junior high school. This problem cannot be solved since the verification depends on the students' information in the INEP data from the Confidentiality Room.

Methodology applied

Considering students with irregular trajectories and the six steps presented in the project design, the project used a machine learning methodology based on the methods known in the literature as Logistic Regression (Baseline) and Xgboost to estimate the probability of a student dropping out before completing elementary and junior high school, using academic records up to the 5th grade.

- Logistic regression: Baseline model generated using the OLS algorithm without considering terms of interaction. The objective was to create a performance benchmark enabling the measurement of real gain by increasing the complexity of the final model, using a more advanced machine learning technique.

- Xgboost: To obtain better performance and mitigate possible **Covariate Shift** problems, we use the **Gradient Boosting Decision Tree** machine learning algorithm. This algorithm is currently state-of-the-art for generating machine learning models on tabular data and naturally handles missing values and categorical variables.



The diagram illustrates the training and testing process used for the machine learning model. Initially, the data is carefully sanitized, forming a 'Clean Base' with training sets for the year 2014, validation for 2015 and test for 2018. This

selection of data suggests a temporal validation methodology, where data are chosen based on specific periods to simulate a realistic application of the model at different time frames.

Subsequently, the model was subjected to a cross-validation procedure, a technique that involves dividing the data into multiple partitions to ensure that the model is tested on multiple subsets of data, thus promoting a comprehensive evaluation of its performance.

In the process, two crucial optimization phases were used: the first is 'Hyperparameter Optimization', which adjusts the model parameters to achieve the best possible performance; the second is 'Feature Engineering and Selection', which identifies and refines the most significant variables that influence the model's predictions.

The 'Performance Check' was performed to evaluate the accuracy and effectiveness of the model, resulting in the calculation of the 'Official PR-AUC'. PR-AUC, or area under the Curve–Precision Recall, is a widely used metric to measure model quality in contexts where classes are unbalanced.

The 'Objective Function: Log-loss', which works as a compass guiding the optimization process, and the 'Error Function: PR-AUC', which calculates the discrepancy between the model's predictions and the expected results, were also used.

Predictor performance analysis

Due to the nature of the problem having a severe class imbalance, it was necessary to choose a metric that best translates the performance of a statistical model in such a situation. As mentioned earlier, the most recommended metric for datasets with extreme unbalance is the **Area Under the Curve x Precision Recall (PR-AUC);**

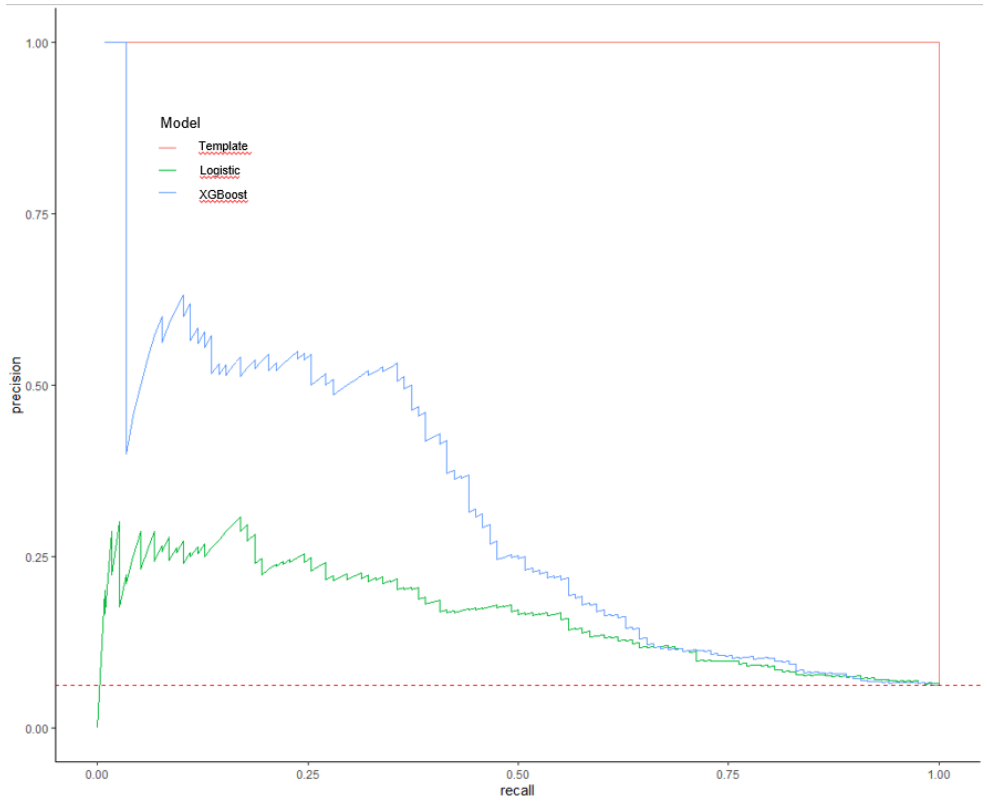
Given an arbitrary cutoff point in the distribution of students according to the probability of dropping out, which defines, for example, a list that will be the target of a specific intervention, the measures that are the basis of the performance analysis are defined as:

a. ***Precision:*** represents the percentage of students who enter the final list who actually belong to the group of students who dropped out;

b. ***Recall:*** represents the percentage of students who actually dropped out, who appear on the final list;

Each pair of Precision and Recall values is always associated with the same probability cutoff point that defines which students make up the final list. This makes it crucial to determine this cutoff point for the practical use of the model, avoiding waste of operational effort during the planned intervention or the non-inclusion of important cases in the final list generated.

The following graph and table present the performance results achieved by the prediction model.

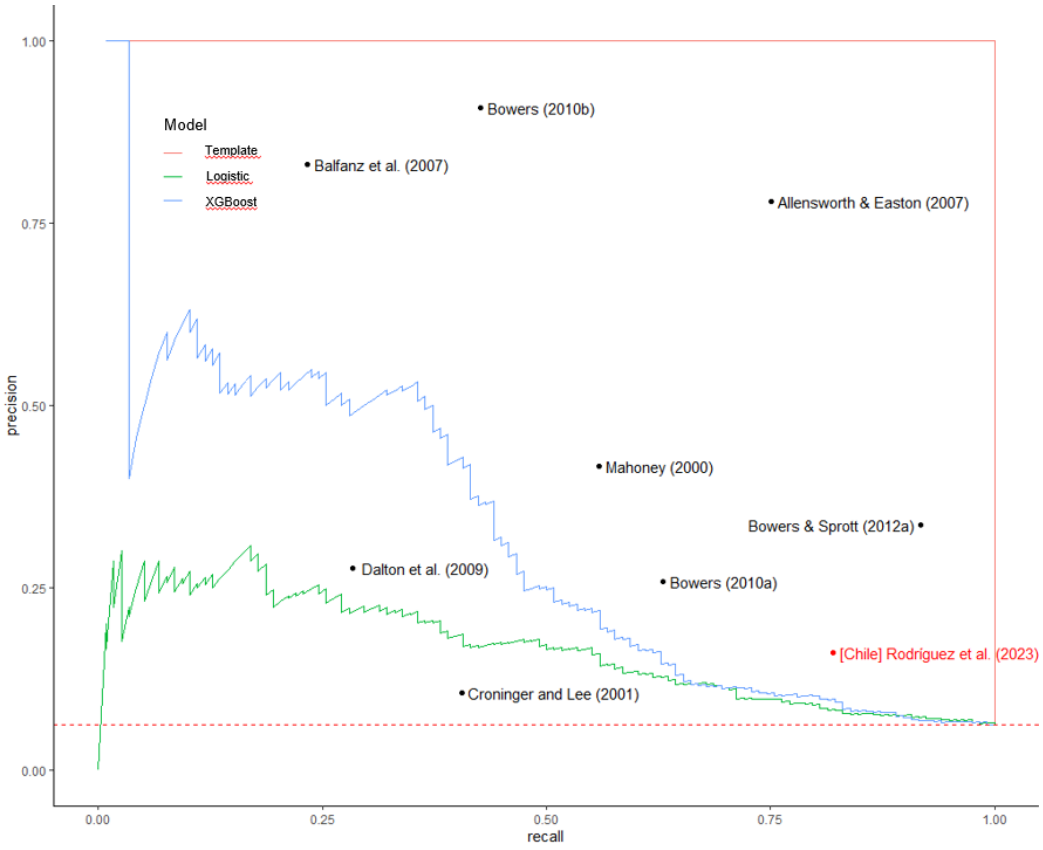


Listing size	List hit rate (precision)	Coverage Rate (Recall)
5%	50%	36%
12%	25%	50%
25%	15%	63%
50%	9%	82%

Increased inclusion error

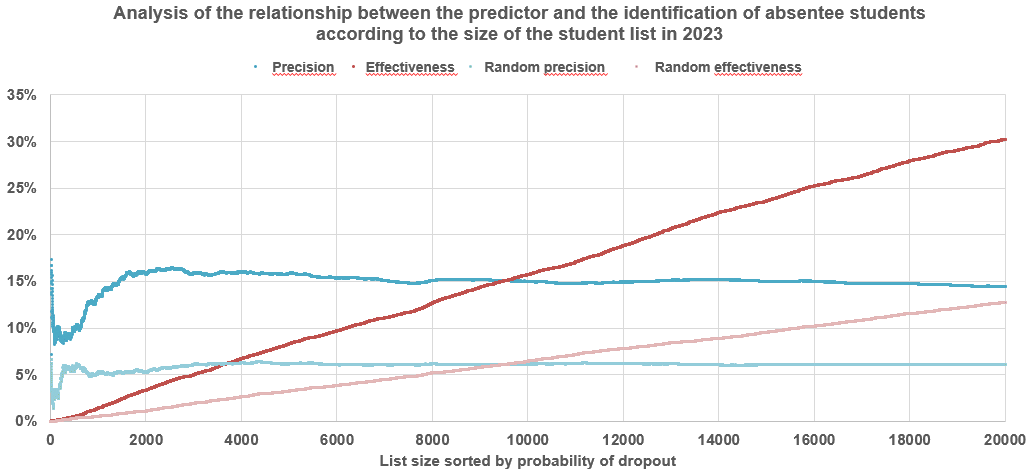
Reduction of exclusion error

Even considering the analysis only among students with irregular trajectories, the prediction model showed a performance of median to low compared to other prediction models. The following graph shows the results of other "similar" prediction models.



It is important to consider that the challenges faced by the other studies regarding the identification and imbalance of the target variable are much less complex than the object of this study. In any case, the question remains whether the model's predictive capacity compensates for the efforts to effectively apply and improve the policy.

An additional analysis was carried out with a list for 2023, comparing students with the highest probability of dropping out according to the predictor with the identification of higher student non-attendance in 2023, since absenteeism is considered to be a measure strongly correlated with dropout. The result of the predictor has a low predictive power, but still superior to the absence of any additional type of information.



Precision: % of students on the list who have low attendance

Effectiveness: % of low attendance students who are identified on the list

Key challenges and lessons learned

- Challenge 1: the dropout event, although very important, is rare in elementary and junior high school and can occur at some point throughout basic education. Attempts to predict high school dropout end up being more successful due to the higher incidence of this event at this stage of the basic cycle.

- Challenge 2: the *covariate shift* problem identified in the administrative data, when the variables are not approximately independent and identically distributed over time, end up restricting the use of available information and limiting the predictive power of the model. This can occur due to the absence of information or the inaccuracy of the records.

- Challenge 3: the difficulty of identifying exactly the information on the occurrence of the dropout due to not having visibility of the student's history when he leaves the network by transfer generates an important information noise for the main variable of interest.

Continuity of studies

- To overcome the challenges and contribute to the main purpose of this study – to develop a tool to support SME programs to combat dropout – we recommend that the project should continue with a proposal to develop a predictor of student absenteeism. This proposal solves challenges 1 and 3 described above and is directly related to the main indicator for coping with school dropout defined by the *Bora Pra Escola* program.

- The absenteeism or non-attendance predictor would aim to identify students with a probability of reaching an absenteeism higher than the minimum acceptable limit in year T, based on information from the academic record of year T-1, and characteristics of the student and of the school. From the result of the predictor, it would be possible to generate a list at the beginning of the school year of the students most likely to reach the undesired level of absenteeism. The list could direct the actions of the program and the efforts of the network to the places where the greatest incidence of the event is expected. As in the previous model tested, there is a commitment to evaluate the performance of the model and verify the real predictive capacity of the model before it is used, including checking the existence of the *Covariate Shift* between the years of the database provided, for the structuring of data necessary for the absenteeism predictor.

- In addition, it would be very useful, as a strategy to prevent dropout, to develop a predictor of failure, which would aim to identify the students most likely to fail in year T, having as reference the information of the school record of year T-1 and the characteristics of the student and the school. A predictor of failure could signal to students who need individualized support and school reinforcement, which can both remedy situations that generate failure by grade and, in extreme cases, failures due to absence and school dropout.

References

Bowers, A.J., Sprott, R. and Taff, S.A., 2012. **Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity**. The High School Journal, pp.77-100.

Chen, T., & Guestrin, C. (2016, August). **Xgboost: A scalable tree boosting system**. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Rodríguez, P., Villanueva, A., Dombrovskaia, L. and Valenzuela, J.P., 2023. **A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile**. Education and Information Technologies, pp.1-47.

[i] Report produced with technical assistance from Oppen Social.