imds
instituto mobilidade e
desenvolvimento social

Instituto Mobilidade e Desenvolvimento Social

Preditor de evasão para os anos finais do ensino fundamental na rede municipal do Rio de Janeiro – Desafios e Aprendizados.[i]

Rio de Janeiro, 20 de dezembro de 2023

Introdução

O Brasil, ao longo das últimas décadas, tem enfrentado desafios persistentes relacionados ao abandono e à evasão escolar no ensino fundamental. Essa problemática, que afeta diretamente o desenvolvimento social e econômico do país, possui raízes multifacetadas, englobando fatores socioeconômicos, pedagógicos e estruturais. Apesar dos esforços governamentais para universalizar o acesso à educação, a retenção de alunos nas escolas permanece como um obstáculo crítico, com impactos que se estendem para além dos muros escolares, refletindo-se no mercado de trabalho e na sociedade como um todo.

A evasão escolar é uma questão complexa que exige uma abordagem multidimensional, incorporando a compreensão das variáveis que influenciam a decisão do aluno de abandonar a escola. Dentre esses fatores, destacam-se a qualidade do ensino, o contexto familiar, questões de infraestrutura e recursos educacionais, bem como o engajamento e a relevância percebida do currículo escolar para a vida do estudante. Um preditor de evasão escolar pode ser uma ferramenta diferenciada na identificação precoce de sinais de risco, permitindo intervenções direcionadas e eficazes que podem alterar trajetórias de vida e fortalecer a educação como um pilar do processo de mobilidade social.

O projeto Preditor de evasão para os anos finais do ensino fundamental na rede municipal do Rio de Janeiro visa atuar como um agente transformador, inserindo- se estrategicamente nesse cenário. A necessidade de intervir desde o ensino fundamental é crucial, uma vez que o evento da evasão por muitas vezes se deve a um processo cumulativo que se inicia desde a primeira infância e acaba por ocorrer em algum momento ao longo da educação básica. O abandono nesse estágio não apenas compromete o acesso à educação, mas também impacta a formação integral dos indivíduos, perpetuando desigualdades e limitando oportunidades futuras. Com o propósito de se tornar mais uma frente de informação para apoiar as ações do programa Bora Pra Escola, o projeto busca antecipar-se a essas trajetórias adversas por meio da implementação de um sistema de alerta nos anos finais, alimentado por

dados pregressos dos estudantes que sinalizem maiores ou menores chances de estes evadirem antes de concluir o ensino básico.

O projeto desenvolvido ao longo de 2022 e 2023 no âmbito do ACT entre o Instituto Mobilidade e Desenvolvimento Social (IMDS) e a Secretaria Municipal de Educação do Rio de Janeiro contou com uma série de desafios. O preditor de evasão alcançou uma performance abaixo da esperada e demanda que novas etapas de pesquisa sejam exploradas para que se torne de fato uma ferramenta diferenciada no apoio ao combate da evasão. Este relatório apresenta todas as etapas de desenvolvimento do preditor, a performance alcançada, os desafios enfrentados e principalmente os aprendizados até aqui, de forma que essas informações sejam aproveitadas para continuidade do projeto.

O Projeto inicial

A proposta inicial consistiu em criar um preditor de probabilidade de evasão, baseado em modelagem estatística. A escolha do 5º ano como ponto de análise estratégica se justifica pela sua relevância no percurso educacional, sendo um período decisivo que antecede a transição para a etapa seguinte. O objetivo do projeto foi prever a chance de um aluno evadir antes de concluir o ensino fundamental.

Como resultado concreto, esperava-se obter uma lista anual de alunos que concluíram o 5º ano, acompanhada da respectiva probabilidade de evasão até o final do Ensino Fundamental.

Para alcançar esse objetivo, seriam utilizados insumos como a trajetória escolar do aluno, características pessoais, escolaridade dos pais e características da escola, disponibilizados anualmente pela SME/RJ. A figura a seguir ilustra a proposta inicial do projeto.

O projeto adotou uma metodologia em seis passos que combina observação do histórico escolar, treinamento de máquina e aplicação do modelo em coortes sucessivas:

- **Passo 1:** Observação do histórico escolar e características dos alunos que já passaram pela rede.
- **Passo 2:** Relacionamento do histórico escolar e características dos alunos nos anos iniciais com o resultado de interesse nos anos finais (evadiu ou não).
- **Passo 3:** Treinamento de máquina para identificar padrões entre as informações observadas nos anos iniciais e o resultado de interesse nos anos finais.
- **Passo 4:** Verificação da performance do modelo treinado aplicando-o na coorte seguinte e avaliando o nível de acerto.
- **Passo 5:** Aplicação do modelo em uma coorte mais distante para realizar o teste final de performance.
- **Passo 6:** Predição (inferência) aplicando os parâmetros corrigidos para gerar uma lista de alunos com suas respectivas probabilidades de evasão.

O preditor

Variáveis utilizadas

A informação central de interesse para o preditor é a probabilidade de evasão antes do término do Ensino Fundamental.

A variável escolhida para representar esse fenômeno foi o registro de abandono como última informação de movimentação conhecida em todos os anos, em acordo com a SME sobre a melhor forma possível de identificá-lo nos registros administrativos disponíveis. Dessa forma, a evasão foi definida como 1 quando o estudante teve a última informação de movimentação na rede como “Abandono” e 0 para todos os demais casos. Essa abordagem mais abrangente permite uma identificação mais precisa dos casos de evasão.

A tabela a seguir apresenta a quantidade de registros analisados e a última situação de movimentação identificada.

Situaçao	Ultima_movimentacao	QTD
Inativo	Abandono	577.284
Inativo	Conclusão de nível de ensino	456.227
Inativo	Renovação de matrícula	365.669
Inativo	Transferência para a própria rede	353.090
Inativo	Transferência para rede particular	262.431
Inativo	Mudança de turma	226.800
Inativo	Remanejamento para própria rede	214.765
Inativo	Transferência para rede pública estadual/federal	202.208
Inativo	Transferência para rede pública de outros municípios	160.961
Inativo	Término do PEJA	144.863
Inativo	Necessidade de trabalhar	80.230
Inativo	Carga inicial do sistema	60.901
Inativo	Transferência na própria rede	42.292
Inativo	Duplicidade de matrícula	23.720
Inativo	Matrícula inicial	20.534
Inativo	Desistência de vaga	16.265
Inativo	Recondução	16.210
Inativo	Remanejamento da própria rede	10.803
Inativo	Transferência da rede particular	10.029
Inativo	Doença ou anomalia grave	6.262
Inativo	Transferência de rede pública estadual/federal	5.529
Inativo	Adequação de matrícula	5.081
Inativo	Falecimento	4.805
Inativo	Regularização de COC de recurso	3.192
Inativo	Transferência da rede pública de outros municípios	2.984
Inativo	Mudança de modalidade de ensino	2.495
Inativo	Reclassificação (Entrada)	1.932
Inativo	Mudança de bloco de PEJA	423
Inativo	Adequação de matrícula	202
Inativo	Reclassificação	1

As variáveis preditoras foram extraídas dos arquivos de turma e avaliações, conforme descrito abaixo:

Sobre o aluno:

- Escolaridade do 1º responsável
- Escolaridade do 2º responsável
- Trajetória regular ou irregular
- Frequência no 5º ano
- Conceito no 5º ano

Sobre a turma:

- Último ano em que frequentou o 5º ano
- Quantidade de mudanças de escola
- Quantidade de reprovações

Sobre a escola:

- Índice de Complexidade da Gestão
- Percentual de alunos em cada nível socioeconômicos (definidos pelo INSE)

Disponibilidade e qualidade dos dados

O processo de acesso e crítica dos dados foi uma etapa que demandou bastante esforço e tempo, especialmente devido aos problemas comuns no trabalho com registros administrativos e os cuidados necessários para atendimento a todas as normas da LGPD.

A SME já vinha realizando um trabalho longo de sistematização e padronização das informações ao longo de 10 anos de histórico dos estudantes e organizando esses dados em um Data Warehouse para que pudesse ser utilizado em pesquisas de apoio à melhoria dos programas e ações da secretaria.

Os dados foram disponibilizados e passaram por um processo de análise crítica e de viabilidade para a modelagem estatística de interesse. A primeira etapa foi verificar a premissa de que as variáveis são aproximadamente iid (independentes e identicamente distribuída) ao longo do tempo, quando comparados os coortes. Quando isso não ocorre, esse fenômeno é chamado de Covariate Shift. O procedimento é composto da seleção de dois subconjuntos de dados que tem 4 anos de diferença e da atribuição da variável target o valor 1 a todos os registros do ano mais recente e 0 aos do ano mais antigo. Em seguida, do treinamento de um modelo de classificação, usando a variável target criada na etapa anterior.

Para mensurar se o modelo treinado é capaz de identificar os registros pertencentes a cada ano usando apenas as variáveis independentes como referência, foi usada a métrica ROC-AUC (area under the receiver operator curve), que varia entre 0.5 e 1.0, onde um valor de 0.5 indica que os anos são indistinguíveis (caso ideal), usando apenas as variáveis independentes, e 1.0 indica que os anos são perfeitamente separáveis (pior caso). Treinando o modelo com os dados recebidos pela SME/RJ e usando todas as variáveis independentes obtivemos um valor da ROC-AUC acima de 0.9, indicando que a distribuição de pelo menos uma variável independente em questão se altera com o passar do tempo.

Esse resultado obtido implica na necessidade de seleção cuidadosa das variáveis independentes a serem usadas no modelo, no intuito de mitigar esse efeito e não obter uma baixa performance do modelo final. Após o processo de seleção de variáveis reexecutando o procedimento descrito anteriormente para inúmeros subconjuntos de variáveis independentes, foi obtido um subconjunto que possuía o menor grau de Covariate Shift possível para o conjunto de dados fornecidos, levando em consideração o tradeoff de perda de performance na predição da evasão. Consequentemente, uma parte das variáveis foi utilizada no modelo e uma outra parte foi descartada por não apresentar bom poder preditivo consistente ao longo do tempo para a predição da evasão. O quadro a seguir sintetiza as variáveis disponibilizadas e aquelas que foram efetivamente utilizadas.

Variáveis disponíveis e utilizadas	Variáveis disponíveis e não utilizadas
Sobre o histórico:	Sobre o aluno:
Última movimentação: se abandono, indica evasão	Gênero
Sobre o aluno:	Naturalidade
Escolaridade do 1º responsável	Raça / Cor
Escolaridade do 2º responsável	Tipo de transporte
Trajetória regular ou irregular	Idade no 5º ano
Sobre avaliações:	Nacionalidade
Frequência no 5º ano	Portador de deficiência
Conceito no 5º ano	Sobre avaliações:
Sobre a turma:	Frequência de 1º a 4º ano
Último ano frequentou = 5º ano	Conceito de 1º a 4º ano
Quantidade de mudança de escola	Sobre a turma:
Quantidade de reprovações	Quantidade de mudança de escola no meio do ano
Sobre a escola:	Quantidade de mudança de turma e turno
Índice de Complexidade da Gestão	Sobre a escola:
Percentual de alunos em cada INSE	INSE

A análise dos dados identificou a necessidade de dividir o arquivo em duas partes, considerando que os estudantes com “trajetórias regulares” em um grupo e os estudantes com “trajetórias irregulares” em outro grupo. As trajetórias irregulares foram definidas pela identificação de alguma interrupção atípica no processo de avanço anual de uma série para outra. O quadro a seguir exemplifica a diferença entre os dois tipos de trajetória:

Exemplos de "trajetória Regular"

1º ano
2º ano
3º ano
4º ano
5º ano
6º ano
7º ano
8º ano
9º ano

Exemplos de "trajetória Irregular"

ano	id_aluno	grupamento
2000	X	Grupo I
2001	X	Grupo I
2001	X	Grupo I
2002	X	1º Ciclo de Formação - Período Inicial
2003	X	1º Ciclo de Formação - Período Intermediário
2003	X	1º Ciclo de Formação - Período Intermediário
2004	X	1º Ciclo de Formação - Período Final
2004	X	1º Ciclo de Formação - Período Final
2005	X	3ª Série
2005	X	3ª Série
2006	X	4ª Série
2006	X	4ª Série
2008	X	3º Ciclo de Formação - Período Inicial
2009	X	8º ano

ano	id_aluno	grupamento
2007	Y	Berçário
2008	Y	Grupo I
2009	Y	Grupo I
2010	Y	1º ano
2011	Y	2º ano
2012	Y	3º ano
2013	Y	Realfabetização 1
2014	Y	4º ano
2015	Y	5º ano
2016	Y	6º ano
2019	Y	Carioca I
2020	Y	Carioca II

A divisão entre os grupos foi necessária pela identificação de que o grupo de estudantes com trajetórias regulares apresentava uma ocorrência muito mais baixa de evasão do que o outro grupo e, portanto, tornava o exercício da predição muito mais desafiador. O quadro abaixo apresenta a diferença entre a ocorrência de evasão nos dois grupos considerando uma amostra aleatória dos estudantes.

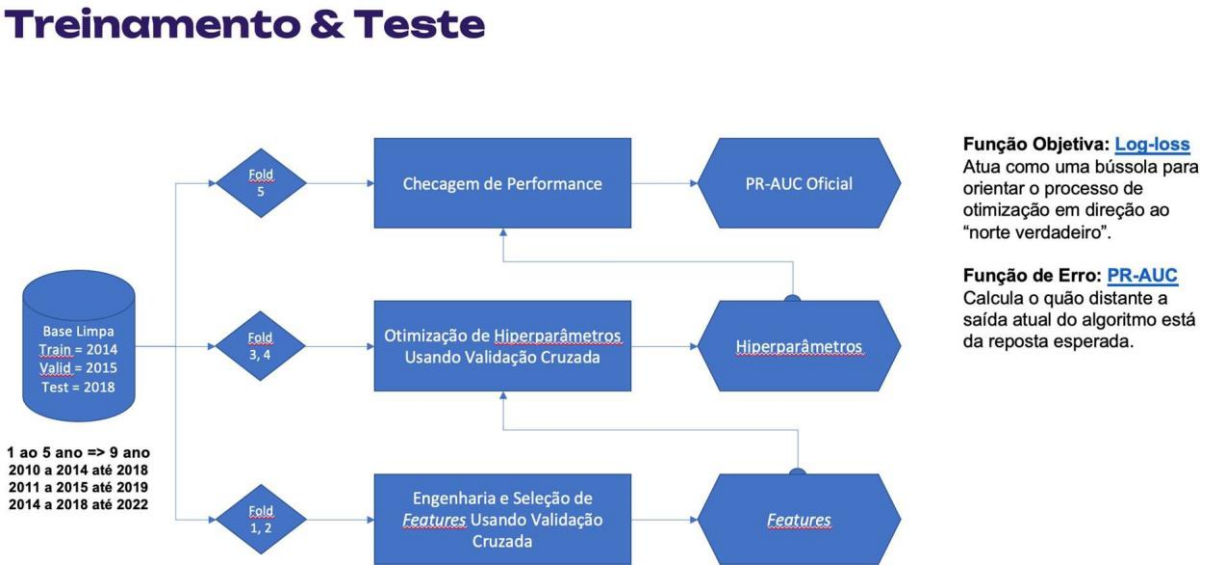
Com "trajetória Regular"		Análise dos estudantes que passaram pelo 5º ano em 2018		Com "trajetória Irregular"		
4.499 estudantes		6.408 estudantes		30% do total	1.909 estudantes	
4.412 não evadiram		6.203 não evadiram			1.791 não evadiram	
87 evadiram	1,9% Taxa de evasão	205 evadiram	3,2% Taxa de evasão	58% do total	118 evadiram	6,2% Taxa de evasão

Um segundo problema identificado na análise dos dados foi a diferença no percentual de evasão identificado nos registros administrativos e nos dados publicados do INEP. Possivelmente, esse problema se deve a falta de informação sobre os estudantes quando estes são transferidos para outras redes antes da conclusão do ensino fundamental. Esse problema não pode ser resolvido uma vez que a verificação depende das informações dos estudantes nos dados do INEP a partir da Sala de Sigilo.

Metodologia aplicada

Considerando os estudantes com trajetórias irregulares e os seis passos apresentados no desenho do projeto, o projeto utilizou uma metodologia de aprendizado de máquina a partir dos métodos conhecidos na literatura como Regressão Logística (Baseline) e Xgboost para estimar a probabilidade de um estudante evadir antes de concluir o ensino fundamental, utilizando as informações históricas até o 5º ano.

- Regressão logística: Modelo baseline gerado utilizando o algoritmo OLS sem levar em consideração termos de interação. Seu objetivo foi criar uma referência de performance possibilitando a mensuração do ganho real ao se incrementar a complexidade do modelo final, utilizando uma técnica mais avançada de aprendizado de máquina.
- Xgboost: Para obtermos uma melhor performance e mitigar possíveis problemas de **Covariate Shift** utilizamos o algoritmo de aprendizado de máquina **Gradient Boosting Decision Tree**. Esse algoritmo atualmente é o estado da arte para gerar modelos de aprendizado de máquina em dados tabulares e lida de forma natural com valores *missing* e variáveis categóricas.



O diagrama ilustra o processo de treinamento e teste utilizado para o modelo de aprendizado de máquina. Inicialmente, os dados são cuidadosamente

higienizados, formando uma 'Base Limpa' com conjuntos de treino do ano de 2014, validação para 2015 e teste para 2018. Esta seleção de dados sugere uma metodologia de validação temporal, onde os dados são escolhidos com base em períodos específicos para simular uma aplicação realista do modelo em diferentes marcos temporais.

Posteriormente, o modelo foi submetido a um procedimento de validação cruzada, uma técnica que envolve a divisão dos dados em várias partições para garantir que o modelo seja testado em vários subconjuntos de dados, promovendo assim uma avaliação abrangente de seu desempenho.

No processo foram utilizadas duas fases cruciais de otimização: a primeira é a 'Otimização de Hiperparâmetros', que ajusta os parâmetros do modelo para alcançar o melhor desempenho possível; a segunda é a 'Engenharia e Seleção de Features', que identifica e refina as variáveis mais significativas que influenciam as previsões do modelo.

A 'Checagem de Performance' foi realizada para avaliar a precisão e a eficácia do modelo, resultando no cálculo do 'PR-AUC Oficial'. O PR-AUC, ou área sob a curva Precision-Recall, é uma métrica muito utilizada para medir a qualidade do modelo em contextos em que as classes são desequilibradas.

Também foram utilizadas a 'Função Objetiva: Log-loss', que funciona como uma bússola norteando o processo de otimização, e a 'Função de Erro: PR-AUC', que calcula a discrepância entre as previsões do modelo e os resultados esperados.

Análise de performance do preditor

Devido à natureza do problema possuir um desbalanceamento grave de classes foi necessário escolher uma métrica que melhor traduz a performance de um modelo estatístico em tal situação. Como mencionado anteriormente, a métrica mais recomendada para conjuntos de dados com desbalanceamento extremo é a **Área Sob a Curva Precision x Recall (PR-AUC)**;

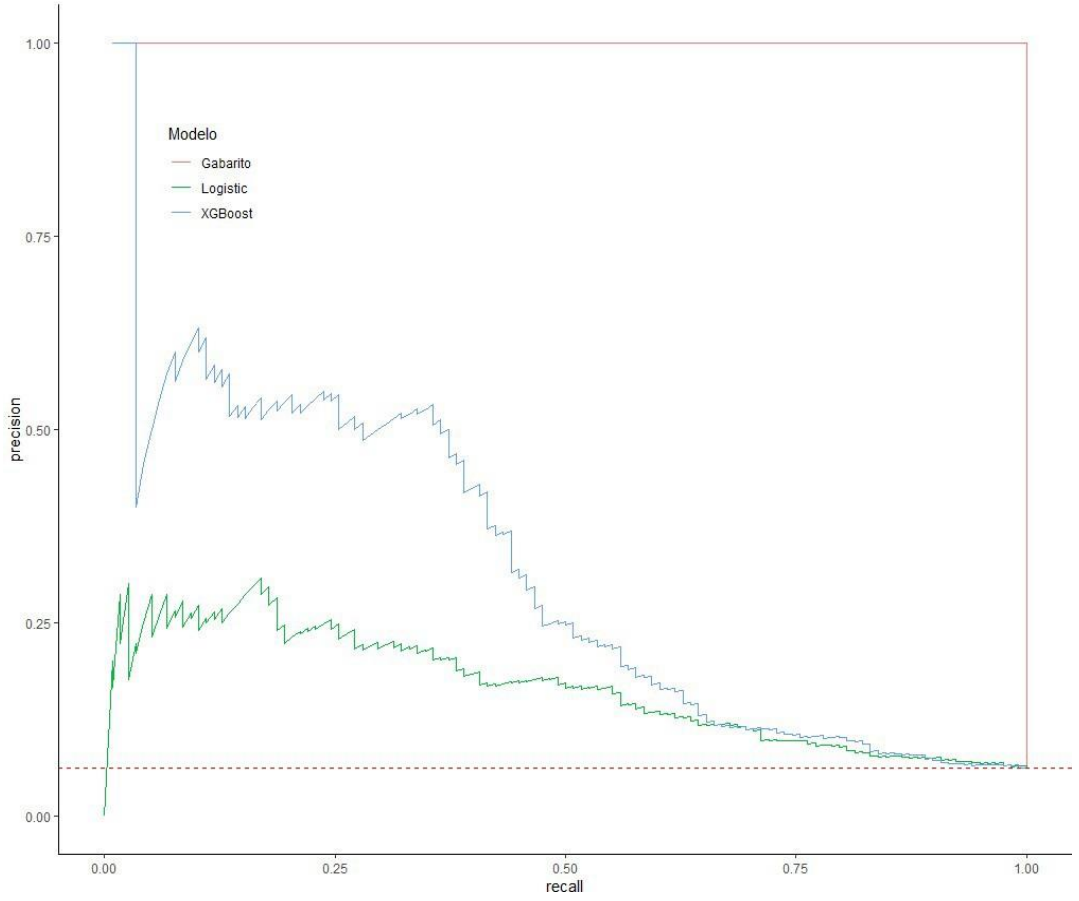
Dado um ponto de corte arbitrário na distribuição dos estudantes segundo a probabilidade de evasão, que define, por exemplo, uma lista que será alvo de uma intervenção específica, as medidas que são a base da análise de performance são definidas como:

a. ***Precision:*** representa o percentual de estudantes que entram na lista final que realmente pertencem ao grupo de estudantes que evadiram;

b. ***Recall:*** representa o percentual dos estudantes que realmente evadiram, que aparecem na lista final;

Cada par de valores Precision e Recall está sempre associado a um mesmo ponto de corte da probabilidade que define quais estudantes compõem a lista final. O que torna crucial a determinação desse ponto de corte para a utilizada prática do modelo, evitando desperdícios de esforço operacional durante a intervenção planejada ou a não inclusão de casos importantes na lista final gerada.

O gráfico e a tabela a seguir apresentam os resultados de performance alcançados pelo modelo de predição.

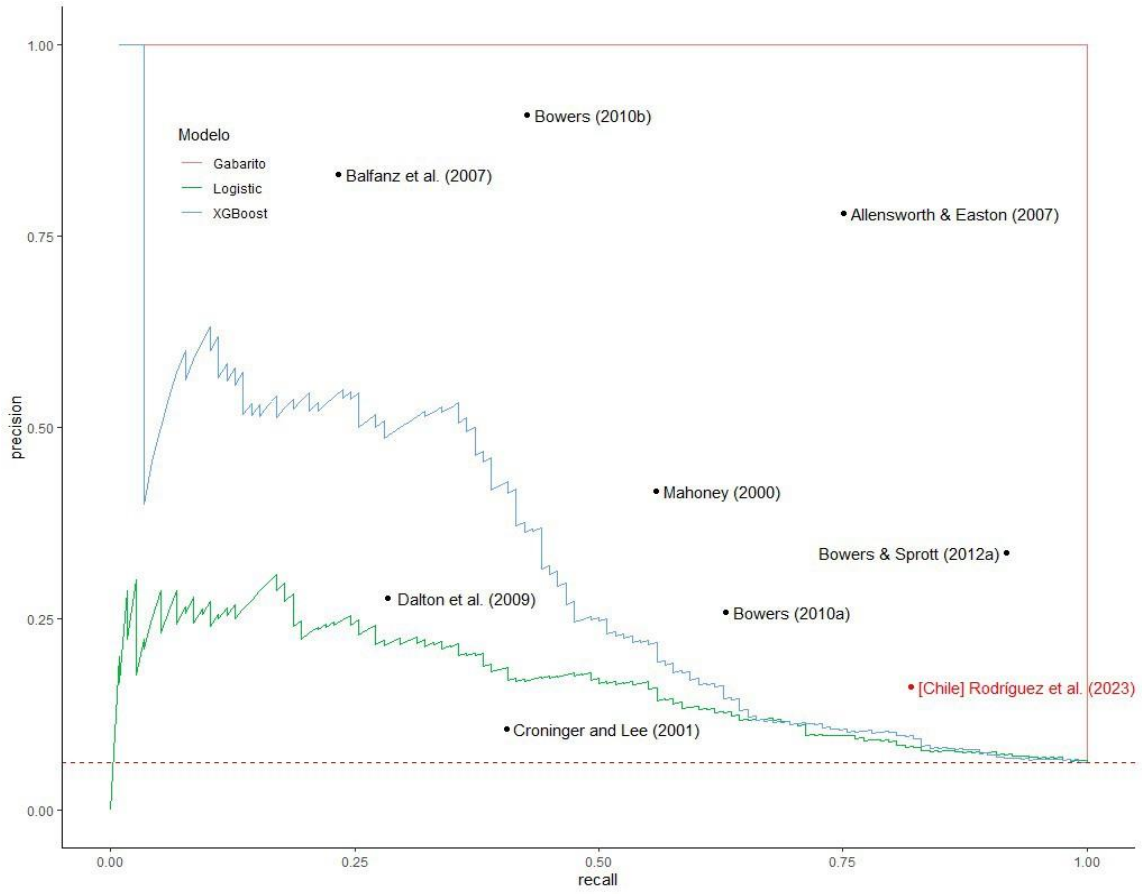


Tamanho da listagem	Taxa de acerto da lista (*precision*)	Taxa de cobertura (*Recall*)
5%	50%	36%
12%	25%	50%
25%	15%	63%
50%	9%	82%

Aumento do erro de inclusão

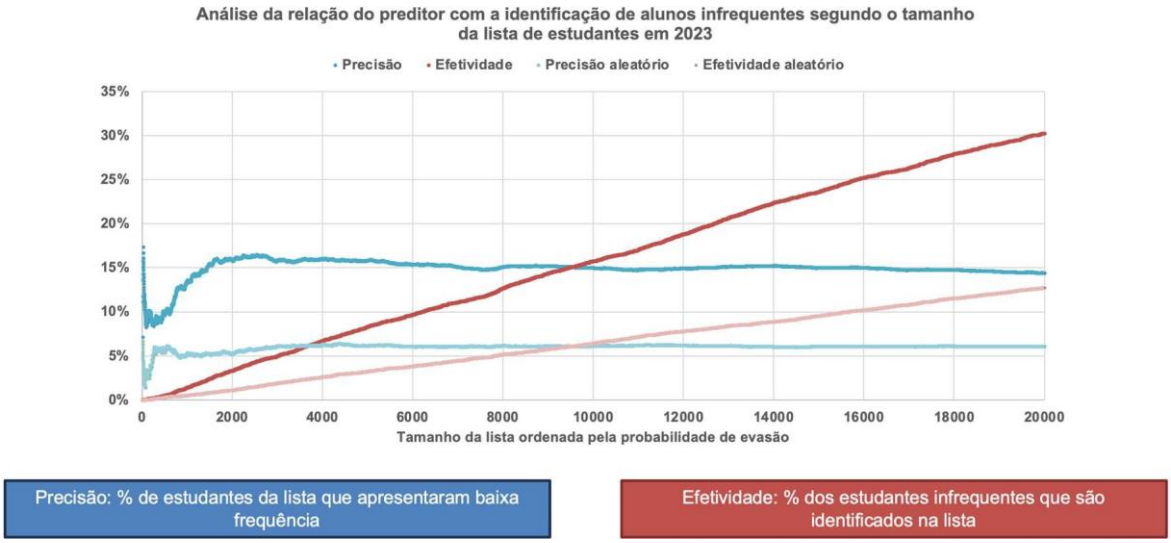
Redução do erro de exclusão

Mesmo considerando a análise apenas entre os estudantes com trajetórias irregulares, o modelo de predição apresentou uma performance de mediana para baixa se comparada a outros modelos de predição. O gráfico a seguir apresenta os resultados de outros modelos de predição “similares”.



É importante considerar que os desafios enfrentados pelos demais estudos quanto a identificação e o desbalanceamento da variável target são bem menos complexos do que o objeto desse estudo. De toda forma, fica a questão se a capacidade de predição do modelo compensa os esforços para aplicação e melhoria da política de forma efetiva.

Uma análise adicional foi realizada com uma lista para 2023, comparando os alunos com maior probabilidade de evasão segundo o preditor com a identificação de maior infrequência dos estudantes em 2023, uma vez que a infrequência é considerada uma medida fortemente correlacionada com a evasão. O resultado do preditor apresenta um baixo poder de predição, mas ainda superior a ausência de qualquer tipo adicional de informação.



Principais desafios e aprendizados

- Desafio 1: o evento da evasão apesar de muito importante, é raro no ensino fundamental e pode ocorrer em algum momento ao longo de toda a educação básica. Tentativas de predizer a evasão no ensino médio acabam por ter maior sucesso devido a maior incidência desse evento nessa etapa do ciclo básico.
- Desafio 2: o problema de *covariate shift* identificado nos dados administrativos, quando as variáveis não são aproximadamente independentes e identicamente distribuídas ao longo do tempo, acabam por restringir o uso de informações disponíveis e limitar o poder preditivo do modelo. Isso pode ocorrer pela ausência de informação ou pela imprecisão dos registros.
- Desafio 3: a dificuldade de identificar exatamente as informações de ocorrência da evasão por não ter visibilidade do histórico do estudante quando este sai da rede por transferência gera um ruído de informação importante para a principal variável de interesse.

Continuidade dos estudos

- Para superar os desafios e contribuir para o principal propósito desse estudo – de desenvolver uma ferramenta para apoiar os programas da SME de combate à evasão – indicamos que o projeto tenha continuidade com uma proposta de desenvolvimento de um preditor de infrequência dos estudantes. Essa proposta resolve os desafios 1 e 3 descritos acima e está diretamente relacionado com o principal indicador de enfrentamento da evasão escolar definido pelo programa Bora Pra Escola.
- O preditor de infrequência teria como objetivo identificar os estudantes com probabilidade de alcançar uma infrequência superior ao limite mínimo aceitável no ano T, a partir de informações do histórico escolar do ano T-1 e de características do aluno e da escola. A partir do resultado do preditor seria possível gerar uma lista logo no início do ano letivo dos estudantes com maior probabilidade de atingir o nível indesejado de infrequência. A lista poderia direcionar as ações do programa e os esforços da rede para os locais onde se espera maior incidência do evento. Da mesma forma como no modelo anterior testado, existe o compromisso de avaliar a performance do modelo e verificar a real capacidade de predição do modelo antes que este seja utilizado. Incluindo a checagem da existência de *Covariate Shift* entre os anos da base fornecida, para a estruturação de dados necessária para o preditor de infrequência.
- Adicionalmente, seria bastante útil, como estratégia de prevenção de evasão, o desenvolvimento de um preditor de reprovação, que teria como objetivo identificar os estudantes com maior probabilidade de reprovar no ano T, tendo como referência as informações do histórico escolar do ano T-1 e de características do aluno e da escola. Um preditor de reprovação poderia sinalizar para alunos que precisam de suporte individualizado e reforço escolar, que pode tanto remediar situações geradoras de reprovação por nota e, em casos extremos, reprovações por falta e evasão escolar.

Referências

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] Relatório produzido com assistência técnica da Oppen Social.