APLICAÇÕES DE MACHINE LEARNING EM ECONOMIA DA EDUCAÇÃO

Túlio Oliveira Prates, Marcos Vinicio Wink Junior e Analucia Vieira Fantin

INTRODUÇÃO

As crescentes taxas de abandono no ensino superior representam um desafio significativo para a educação no Brasil. Estas implicam tanto em perda de oportunidades, dado que indivíduos deixam de obter empregos mais bem remunerados, como também na má alocação de recursos públicos, pois parte do orçamento é destinado a capital humano que não se concretiza. Diante deste cenário, o projeto tem por objetivo identificar determinantes do abandono, utilizando ferramentas de predição baseadas em *Machine Learning*, capazes capturar relações complexas não detectadas por métodos estatísticos tradicionais.

Grande parte da literatura sobre as causas do abandono tem se baseado em dados de alta granularidade. Embora válida, esta abordagem limita a inclusão de variáveis de caráter socioeconômico no processo de modelagem. Além disso, tende a restringir a análise ao nível local, dificultando a generalização de resultados. Isto posto, o modelo proposto tem por objetivo ser de escopo abrangente, com ênfase nos aspectos socioeconômicos do abandono.

DESENVOLVIMENTO

O projeto utilizou-se da abordagem de *Machine Learning* (ML), em especial de um algoritmo de classificação de Florestas Aleatórias, visando desenvolver um modelo capaz de prever, dentre um grupo de ex-estudantes, aqueles que teriam abandonado cursos de graduação.

Tomou-se como base de dados uma agregação do complemento de educação da Pesquisa Nacional de Amostra de Domicílios Contínua (PNADC) dos anos de 2022 a 2024. A PNADC tem como vantagem sua riqueza em informação, tornando-a propícia para a metodologia de ML. Os dados foram filtrados de modo que o conjunto final continha apenas ex-estudantes de ensino superior, com idades entre 18 e 24 anos. O resultado foi uma base com 7621 indivíduos, dos quais 69,81% concluíram o ensino superior e 2301 30,19% o abandonaram. Esta foi dívida entre conjuntos de treino (80% das observações) e teste (20% das observações).

Entre as variáveis utilizadas na construção do modelo, mencionam-se também algumas especificações. Um fato estabelecido na literatura é a importância da escolaridade dos pais na determinação do abandono (AINA et al., 2021). Este tipo de informação não é explicitamente informado pela PNADC e pode ser de difícil identificação. Dito isto, é possível computar uma *proxy* na forma da maior escolaridade do restante de um domicílio. Isto é, a maior escolaridade encontrada entre residentes que não são o próprio indivíduo.

RESULTADOS

Foram tomadas algumas métricas como base para a avaliação da performance do modelo. No que se diz respeito ao processo preditivo, o modelo apresentou uma acurácia de 0,72. Outra métrica de demasiada importância é o chamado *f1-scor*e, uma média harmônica entre as métricas de *precision* e *recall*. O *f1-score* obtido sobre a classe de estudantes que abandonaram foi de 0,57. A performance do modelo foi similar a de outros propostos na literatura recente sobre o tema (MATTI VAARMA; LI, 2024).

Ainda que modelos de ML sejam muitas vezes capazes de realizar predições com performance superior à de modelos estatísticos, a abordagem de ML é tipicamente a-teórica, produzindo resultados de difícil interpretação. Isto posto, tomou-se por objetivo utilizar da abordagem interpretativa do SHAP (*Shapely Additive Explanations)* (LUNDBERG; LEE,

2017), de modo a mensurar a importância de cada variável para o abandono, assim como visualizar as complexas relações que caracterizam o processo de conclusão do ensino superior.

Primeiramente, a abordagem do SHAP pode ser utilizada como forma de mensurar a importância de cada variável na predição do abandono, como demonstrado na Figura 1. Também é possível visualizar as complexas relações de cada característica no processo de predição do abandono, como demonstrado na Figura 2. Neste gráfico, o gradiente no eixo vertical indica se uma observação tem um valor alto ou baixo para cada variável, enquanto o eixo horizontal mede o impacto de cada uma sobre o processo de predição.

CONSIDERAÇÕES FINAIS

As relações complexas demonstradas por métodos interpretativos, como o SHAP, se demonstram potencialmente úteis para a elaboração de políticas públicas. Dito isso, vale ressaltar potenciais limitações. A escolha pela PNADC como base de dados implica na impossibilidade da utilização de informações sobre performance acadêmica de estudantes, estabelecida na literatura como importante determinante para o abandono.

Há também um problema de causalidade reversa na metodologia. Dado que o conjunto de dados é composto por ex-estudantes, a situação do estudante no momento que abandonou ou concluiu o ensino superior pode não ser a mesma ao responder a pesquisa da PNADC. Este problema é minimizado ao limitarmos a análise àqueles entre 18 e 24 anos, garantindo que o tempo entre o abandono e a pesquisa é pequeno.

Palavras-chave: Machine Learning, Random Forests, Abandono

ILUSTRAÇÕES

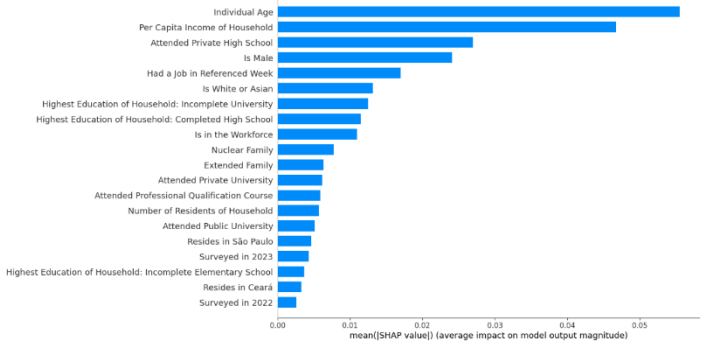

Figura 1. *Mean Absolute SHAP Value*

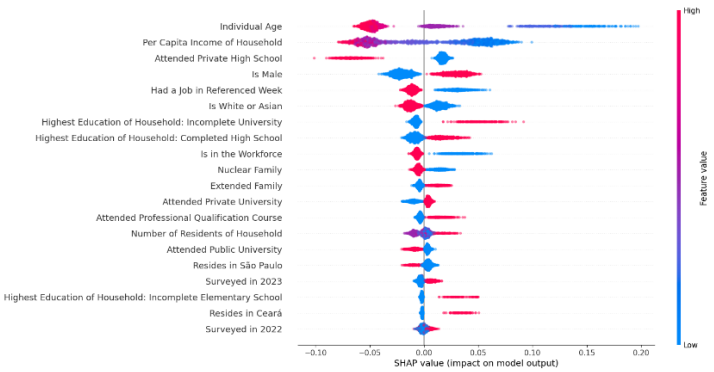

Figura 2. *SHAP Beeswarm Plot*

REFERÊNCIAS BIBLIOGRÁFICAS

AINA, C. et al. The Determinants of University dropout: a Review of the socio-economic Literature. **Socio-Economic Planning Sciences**, v. 79, n. 101102, p. 101102, jun. 2021.

MATTI VAARMA; LI, H. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. **Technology in Society**, v. 76, p. 102474–102474, 1 fev. 2024.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. v. 30, p. 4768–4777, 4 dez. 2017.

DADOS CADASTRAIS

BOLSISTA: Túlio Oliveira Prates
MODALIDADE DE BOLSA: PROBIC
VIGÊNCIA: 09/2024 a 08/2025 – Total: 12 meses
ORIENTADOR: Marcos Vinicio Wink Junior
CENTRO DE ENSINO: ESAG
DEPARTAMENTO: Departamento de Ciências Econômicas
ÁREAS DE CONHECIMENTO: Ciências Sociais Aplicadas, Economia, Economia da Educação
TÍTULO DO PROJETO DE PESQUISA: Políticas públicas e o desempenho escolar na educação básica
Nº PROTOCOLO DO PROJETO DE PESQUISA: NPP78-2021