



Interpretabilidade de modelos aplicados aos dados do Enem

Davi Keglevich Neiva¹

ICMC-USP

Paulino Ribeiro Villas Boas²

Embrapa

1 Introdução

O desenvolvimento de novas tecnologias nos últimos anos, tanto de hardware quanto de *software*, viabilizou a aplicação de técnicas mais sofisticadas de modelagem, como as de aprendizado de máquina. Focadas principalmente em aumentar o poder preditivo, essas técnicas muitas vezes se baseiam em abordagens não paramétricas e não lineares, que resultam em modelos mais precisos e menos interpretáveis [3].

É diante desse cenário que a interpretabilidade de um modelo pode se tornar um fator tão importante quanto a precisão de suas predições [2]. Por esta razão, metodologias como *Shap Values* foram desenvolvidas para apresentar relações entre as variáveis preditivas e a variável resposta, que muitas vezes não são tão evidentes.

Assim, o objetivo deste trabalho é elencar quais são as informações mais relevantes para um modelo aplicado aos dados socioeconômicos do Exame Nacional do Ensino Médio (Enem), bem como identificar como elas impactam o seu funcionamento através dos *Shap Values*.

2 Banco de dados do Enem

Para exemplificar a aplicação da metodologia do *Shap Values* vamos utilizar os microdados do ENEM (Exame Nacional do Ensino Médio), disponibilizados pelo governo federal no site do ministério da educação. [1]

Nesses dados constam os registros anonimizados de 5,783,109 pessoas que estavam inscritas na edição do Exame de 2020. São informações como: pontuação obtida em cada uma das 5 provas do exame (Redação, Ciências da Natureza, Linguagens e códigos, Ciências Humanas e Matemática), dados do participante (Faixa etária, cor/raça, sexo e etc.), características da escola do participante e o questionário socio econômico (preenchimento de 25 questões).

¹davi_neiva@usp.br

²paulino.villas-boas@embrapa.br

3 Modelo desenvolvido

No Enem, o participante pode obter um diploma de ensino médio, caso consiga nota mínima de 500 pontos na redação e 450 pontos nas demais provas. Dos inscritos que não faltaram no exame, apenas 38,85% conseguiram essas notas mínimas.

O modelo proposto neste trabalho teve como objetivo encontrar quais são as características (disponíveis nos dados do participante, da escola e do formulário socioeconômico) que aumentam ou diminuem a propensão da pessoa alcançar a nota mínima para a obtenção do diploma de ensino médio. As alternativas do formulário foram criadas de forma que valores mais altos representem uma melhor condição econômica; para os dados categóricos do participante, foram criadas variáveis *dummies*.

Para o desenvolvimento do modelo foi feita uma seleção aleatória em que 70% dos registros foram utilizados no treinamento e 30% na validação. A técnica utilizada foi a de *Boosting* através do algoritmo *LightGbm*, que apresenta grande capacidade de extração de relações complexas entre as variáveis.

4 Resultados e discussões

Uma das formas de visualizar os *Shap Values* que foram calculados é através do *Summary plot*, onde as variáveis são ordenadas de acordo com aquelas que mais impactaram na discriminação do modelo (Figura 1). Cada ponto indica a contribuição de uma variável para uma pessoa específica, o valor do *Shap Value* é indicado pela posição do gráfico (quanto mais à direita maior a contribuição da variável na classificação em questão) e o valor da respectiva variável é indicado pela cor.

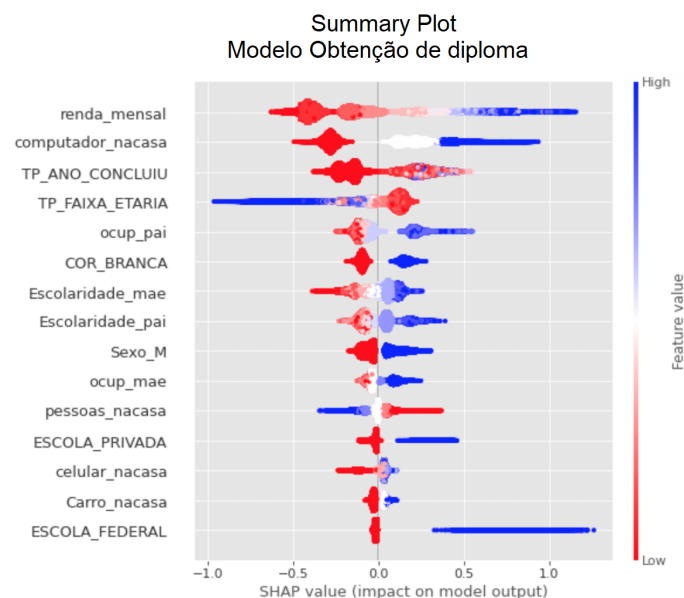


Figura 1: Summary plot dos Shap Values do modelo.

No caso do modelo que foi desenvolvido, cada ponto indica o impacto da respectiva variável

para um participante do Enem ser mais ou menos propenso a obter o diploma. Como esses são modelos de classificação que utilizaram a função logito a escala apresentada para os *Shap Values* é a *Log Odds*. Portanto, quanto mais à direita (maior o valor do *Shap Value*) maior a contribuição da variável para a probabilidade do participante obter nota mínima de diploma no exame ser maior (e quanto mais à esquerda maior a contribuição da variável para uma menor probabilidade).

Podemos observar que a variável que, no geral, mais impactou foi a de renda mensal. Vale notar que pontos em azul mais escuro indicam um maior valor na variável, no caso uma maior renda do participante. A preponderância das cores evidenciam que pessoas com rendas mensais maiores tendem a ter uma propensão maior de alcançar a nota mínima para obtenção do diploma, enquanto que pessoas com baixa renda têm menor chance de concluir o ensino médio através dessa prova.

Outro ponto interessante de notar é que a variável *computador_nacasa* é a segunda variável que mais impactou nas classificações desse modelo. Devido à pandemia e ao distanciamento social desde março de 2020, o exame do Enem 2020 foi extraordinariamente aplicado em janeiro de 2021. Em vista disso, faz sentido o fato da pessoa ter um ou mais computadores influenciar muito no seu desempenho.

No gráfico da Figura 1, também podemos observar alguns comportamentos não lineares. A variável *TP_ANO_CONCLUIU*, que indica o tempo após o término do ensino médio, aparece no gráfico da Figura 1 com uma maior variação de cores à direita. Essa variação indica que dois participantes com tempos de término do ensino médio muito diferentes podem ser impactados por essa variável de forma muito similar. Isso se deve à combinação das características. Para um grupo específico, um longo tempo após o ensino médio aumenta suas chances de obter diploma, enquanto que para outro grupo, as chances diminuem.

Os *Shap Values* também evidenciam que participantes que têm pais com uma maior escolaridade e com ocupações que usualmente são mais valorizadas também vão, no geral, apresentar maiores chances de obter esse diploma. Além disso, é apontado que o sexo e a cor também impactam no desempenho dos participantes, homens e pessoas brancas têm propensão de notas maiores que os demais no modelo devido a essas características. Observamos também que a variável *ESCOLA_FEDERAL*, apesar de ser uma das últimas variáveis (provavelmente porque o número de alunos na rede federal no ensino médio seja bem menor) impacta consideravelmente de forma positiva em quem estuda nessas escolas.

Além dessa visão geral apresentada pelo *Summary_plot* também é possível apresentar uma explicabilidade local. Ou seja, para cada participante do Enem podemos visualizar o impacto de suas características na saída do modelo. Na Figura 2 estão dois exemplos apresentados através do *force_plot*.

Os gráficos da Figura 2 já estão na escala de probabilidade e para o participante 1, indicado no primeiro gráfico, a probabilidade dele alcançar uma nota maior do que a mínima para obter o diploma é alta (82%). O gráfico indica o *base_value* que é o “ponto de partida” no modelo de todas as pessoas e a saída do modelo é alterada conforme as características de cada uma delas. As características apontadas na esquerda em azul aumentam a propensão de obter o diploma, enquanto as da direita diminuem. Observamos que esse participante tem poucas características que diminuem a nota. Apesar do seu pai ter uma ocupação que, geralmente, não é tão valorizada ele possui computador em sua casa, tem acesso a educação da rede federal, é branco e sua mãe completou o ensino médio.

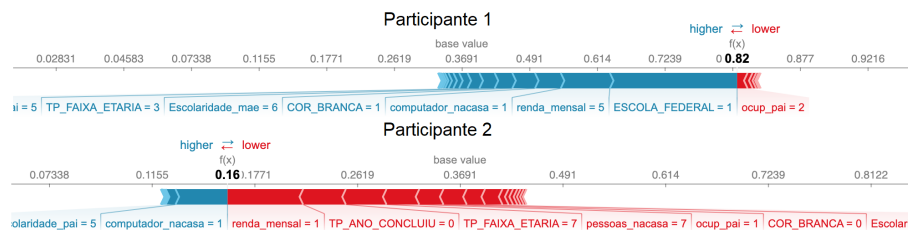


Figura 2: Exemplo modelo diploma.

Já para o participante 2 o modelo indica baixa probabilidade de atingir uma nota acima do mínimo para obter diploma no exame (Figura 2). Mesmo possuindo computador na casa e seu pai ter o ensino médio completo, este participante possui características que impactam negativamente a sua propensão a obter o diploma: i) renda mensal baixa na família, ii) idade 22 anos (indicativo de longo tempo após a conclusão do ensino médio), iii) muitas pessoas vivendo na mesma casa (o que pode gerar concorrência no uso do computador), iv) a ocupação pouco valorizada do pai e v) não ser branco.

5 Conclusões

A utilização de metodologias como *Shap Values* permite uma maior compreensão do funcionamento de um modelo, uma vez que é revelado como cada característica influencia na probabilidade estimada. Principalmente em modelos de *machine learning*, como os de *boosting*, a contribuição de cada característica na probabilidade é condicionada às demais e através dessa metodologia podemos visualizar essas contribuições de uma maneira global com o *Summary plot* e local com o *force plot*.

No caso do Enem é apontado que renda é a característica mais marcante para contribuir na probabilidade de um participante alcançar uma nota mínima. Além disso temos características associadas, como disponibilidade de computador, idade, escolaridade dos pais e cor da pele. Por fim, a partir dessa metodologia podemos mensurar o impacto de cada uma dessas características, podendo nos auxiliar em pesquisas futuras.

Referências

- [1] <https://www.gov.br/inep/pt-br/acao-a-informacao/dados-abertos/microdados/enem>
- [2] LUNDBERG, S.; LEE, S.-I. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017
- [3] MOLNAR, C.; CASALICCHIO, G.; BISCHL, B. Interpretable machine learning—a brief history, state-of-the-art and challenges. arXiv preprint arXiv:2010.09337, 2020