Técnicas de interpretabilidade: estado da arte

"De longe, o maior perigo da inteligência artificial é o fato de as pessoas concluírem cedo demais que a entendem."
Eliezer Yudkowsky[40]

Conceito

A comunidade científica[41,42] propõe várias definições de "interpretabilidade" e "explicabilidade" de um modelo e tende a fazer uma certa distinção entre elas, embora, na prática, esses conceitos sejam frequentemente usados de forma intercambiável. Em termos gerais, a interpretabilidade estaria ligada à capacidade de explicar a um ser humano os resultados de um modelo (sua relação de causa e efeito), enquanto a explicabilidade está associada à compreensão da lógica interna do algoritmo, como ele é projetado e treinado e as etapas envolvidas na tomada de decisões para chegar a um determinado resultado.

Algumas definições acadêmicas a esse respeito são:

- Interpretabilidade é a capacidade de explicar ou apresentar em termos compreensíveis para um ser humano[3].
- Interpretabilidade é o grau em que um ser humano pode entender a causa de uma decisão[44].
- A explicabilidade do resultado de um modelo é a descrição de como o resultado do modelo foi produzido[45].
- Explicabilidade é o grau em que a mecânica interna de um sistema de machine learning pode ser explicada em termos humanos[46].

A necessidade de explicabilidade e interpretabilidade dos modelos favoreceu o surgimento de técnicas cada vez mais sofisticadas para a interpretabilidade local e global dos resultados dos modelos, e a situação atual é de certa padronização e convergência no uso de determinadas técnicas (por exemplo, PDP, LIME ou SHAP).

Ao mesmo tempo, essas técnicas não resolvem completamente o problema da interpretabilidade e, em determinadas circunstâncias, podem gerar resultados contraditórios ou tendenciosos, que coexistem com outros fatores que podem afetar a interpretabilidade do modelo, como:

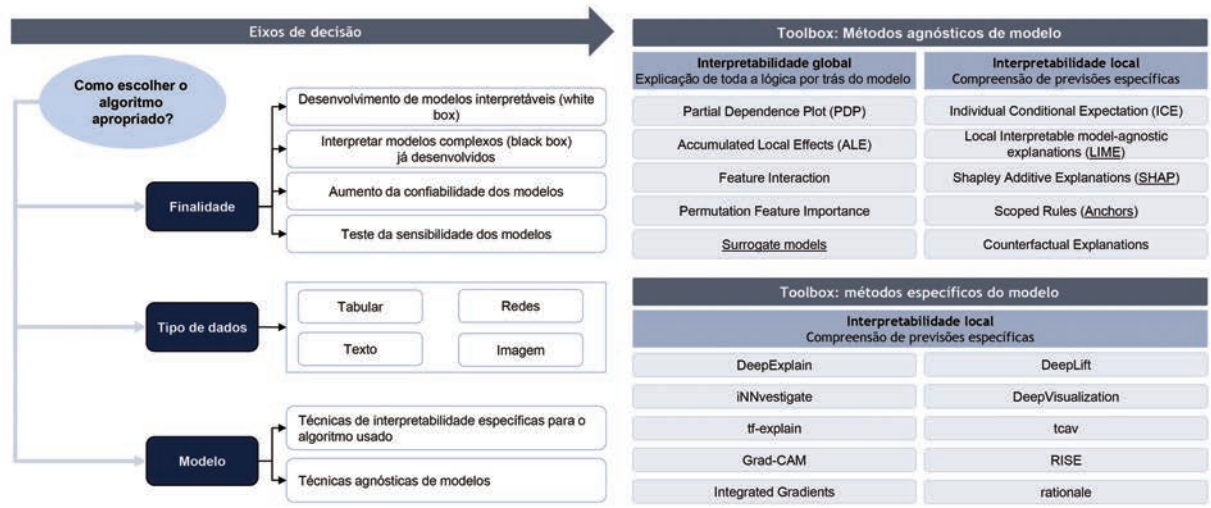
- A reprodutibilidade dos resultados, o processo de treinamento e implementação do modelo[47],a consistência de suas previsões e a explicação da sequência de previsões mais prováveis.
- Potencial de viés[48] nos dados de entrada.
- Imparcialidade (*fairness*)[49].
- Precisão da explicação[50].
- Solidez conceitual do modelo[51].

Para superar várias dessas dificuldades, alguns pesquisadores[52] estão desenvolvendo abordagens alternativas para melhorar a interpretabilidade dos modelos de AI, concentrando-se principalmente no desenvolvimento de modelos inerentemente interpretáveis ("caixas brancas").

Esta seção descreve as principais técnicas de interpretabilidade que são consideradas padrão no setor, bem como o estado da arte no desenvolvimento de caixas brancas.

[40] Eliezer Shlomo Yudkowsky (nascido em 1979), pesquisador e escritor americano especializado em teoria da decisão e inteligência artificial, conhecido por popularizar a ideia de Inteligência Artificial Amigável e defender a Singularidade.
[41] Gall, R. (2018). Editor da Thoughtworks e da The New Stack.
[42] Broniatowsky, D. (2021). Professor Associado, Departamento de Gestão de Engenharia e Engenharia de Sistemas, Universidade George Washington.
[43] Doshi-Velez, F., et al. (2017). Professor de Ciência da Computação na Escola Paulson de Engenharia e Ciências Aplicadas, Universidade de Harvard.
[44] Miller, T. (2019). Professor da Escola de Computação e Sistemas de Informação da Universidade de Melbourne.
[45] Broniatowsky D. (2021).
[46] Gall, R. (2018).
[47] Cientista do Escritório Federal Alemão de Segurança da Informação.
[48] Zhou, N., et al. (2021). Analista financeiro sênior da Wells Fargo.
[49] Ibid.
[50] Jonathon Phillips et al. (2021). Professor de Ciência da Computação e Engenharia, Instituto Nacional de Normas e Tecnologia (NIST).
[51] Sudjianto, A., et al. (2021).
[52] Íbid.

Figura 6. Visão geral das técnicas de interpretabilidade.



Técnicas de interpretabilidade mais comuns

As técnicas de interpretabilidade mais comumente usadas podem ser agrupadas de acordo com sua abordagem[53]: interpretabilidade post-hoc e modelos inerentemente interpretáveis. Há também estratégias complementares para melhorar a compreensão do modelo.

Interpretabilidade post-hoc

As técnicas de interpretabilidade post-hoc, ou interpretabilidade de modelos black box, concentram-se em explicar a saída de modelos treinados com base nas informações fornecidas pelos pesos atribuídos a cada variável de entrada e nos resultados dos modelos. Essas técnicas são úteis para entender os resultados do modelo, embora não forneçam informações sobre o processo de treinamento nem expliquem a lógica interna do algoritmo.

Elas geralmente são divididas em técnicas de interpretabilidade global e local, com referência ao fato de a técnica explicar todo o modelo como um todo ou apenas os resultados em um subconjunto de observações ou dados.

As técnicas de interpretabilidade post-hoc mais comuns são as seguintes (para um inventário mais abrangente, consulte a Fig. 6):

- **PDP** (Partial Dependence Plots, curvas de influência da variável). Essa técnica permite visualizar a influência de cada variável individual no resultado do modelo, excluindo todas as outras variáveis.

- **LIME** (Local Interpretable Model-agnostic Explanations). Essa técnica permite a explicação dos resultados em nível local, ou seja, a explicação dos resultados de uma instância específica com base em informações de outros casos semelhantes.

- **SHAP** (SHapley Additive exPlanations). Essa técnica permite a explicação local e global dos resultados de um modelo, ou seja, a explicação da influência de cada variável nas observações do modelo e a importância de cada variável nos resultados gerais do modelo.

- **Anchors**. Consiste na busca de regras de decisão que expliquem o resultado.

Modelos inerentemente interpretáveis

A interpretabilidade inerente, ou interpretabilidade por modelos white box, concentra-se no desenvolvimento de modelos que são interpretáveis por design ou que podem ser interpretados por construção, por meio de um conjunto de condições que dependem do tipo de modelo (por exemplo, redes neurais[54], em particular ReLu[55], e modelos baseados em árvores[56], entre outros).

Esses modelos permitem uma explicação da lógica interna do algoritmo e da sequência de etapas realizadas para chegar a um resultado específico e, portanto, permitem uma melhor compreensão dos resultados, embora sua aplicabilidade em problemas complexos possa ser mais limitada, dependendo do tipo de algoritmo utilizado.

Estratégias complementares

O uso de estratégias que contribuem para a interpretabilidade dos modelos também pode ser mencionado, como a simplificação do modelo para facilitar sua interpretação, o uso de variáveis com sentido comercial, a análise dos dados para identificar vieses ou falta de imparcialidade (fairness) nas entradas que dificultem a explicabilidade, ou a análise da reprodutibilidade do desenvolvimento do modelo ou de sua implementação, entre outros.

[53] iDanae (2022).
[54] Yang, Z., et al. (2019). Departamento de Estatística e Ciências Atuariais, Universidade de Hong Kong.
[55] Sudjianto. A., et al. (2011).
[56] Sudjianto. A., et al. (2021).

Interpretabilidade post-hoc

1. PDP

Os gráficos PDP[57] (*Partial Dependence Plots,* Gráficos de Dependência Parcial) mostram como a previsão de um modelo AI varia em função de uma ou duas variáveis independentes na previsão, ou seja, o efeito marginal dos preditores. Assim, eles permitem avaliar o tipo de relação entre as variáveis independentes e dependentes.

Sinteticamente:

- Os PDPs mostram graficamente em uma curva a variação média da previsão.
- Essa variação média é obtida variando um preditor para todas as observações no conjunto de dados e, em seguida, obtendo o impacto médio na previsão.
- Uma variante dos PDPs são os gráficos ICE[58] *(Individual Conditional Expectation,* Expectativa Condicional Individual), que mostram de forma semelhante como uma previsão varia para cada observação específica, se um dos preditores do modelo variar, mantendo todos os outros preditores constantes.

2. LIME

LIME[59] *(Local Interpretable Model-agnostic Explanations)* é um método local que testa como as previsões de um modelo variam quando os dados de entrada são perturbados. Para fazer isso, o LIME aplica as seguintes etapas:

- Gerar dados sintéticos em torno da instância de dados de entrada: o LIME toma como ponto de partida uma única previsão e os dados de entrada que a geraram, e gera novos dados de entrada perturbando essa observação, obtendo as previsões correspondentes pelo modelo de AI.
- Treinar um modelo simples em dados sintéticos: o conjunto de dados resultante, composto pelos dados de entrada perturbados e pelas previsões geradas pelo modelo, é usado para treinar um modelo que seja interpretável (por exemplo, modelos lineares, árvores de decisão).
- Explicar as previsões do modelo simples em termos dos dados originais: a importância de cada variável na previsão é obtida, por exemplo, em termos de seus coeficientes na regressão e seu sinal correspondente.

[57] Friedman, J. H. (2001).). Professor do Departamento de Estatística da Universidade de Stanford.
[58] Goldstein, A., et al. (2015). Professor do Departamento de Estatística, The Wharton School, Universidade da Pensilvânia.
[59] Ribeiro, M. T., et al. (2016). Pesquisador da Microsoft Research no grupo de Sistemas Adaptativos e Interação e Professor Adjunto da Universidade de Washington.

Caso de uso: concessão de empréstimos no setor bancário. Uso do PDP.

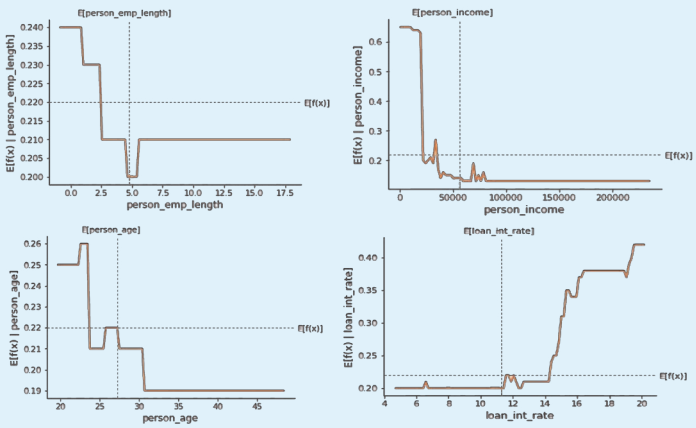
Os PDPs podem ser aplicados a um caso de uso muito comum no setor bancário: a pontuação dos clientes durante o processo de concessão de empréstimo para determinar a probabilidade de inadimplência. Neste exemplo, foi usada uma carteira anônima de empréstimos imobiliários com informações sobre sua atividade nos primeiros três anos.

Foi usado um XGBoost, que é um modelo de árvore não aditivo, um recurso que pode dificultar a explicação. As variáveis usadas pelo modelo durante o treinamento incluem o valor do empréstimo, sua finalidade, o status de propriedade do mutuário, os anos de trabalho em seu emprego atual e a taxa de juros, entre outros.

Nesse contexto, uma área de negócios pode pedir para entender por que o modelo atribui uma determinada probabilidade de inadimplência a um determinado cliente.

Um gráfico PDP mostra a explicação que seria obtida em nível global das variáveis mais envolvidas no resultado e que permitiria ver o impacto que diferentes intervalos dessa variável têm sobre a previsão do modelo (Fig. 7).

Figura 7. PDP para as variáveis "anos de emprego" (em anos), "salário" (euros por ano), "idade" (anos) e "taxa de juros" (vezes um). O eixo X representa a própria variável em estudo, e o eixo Y representa o impacto que diferentes intervalos de cada variável têm sobre a previsão do modelo.



- Calcular a explicabilidade: a porcentagem de explicabilidade pelo LIME é equivalente ao coeficiente de ajuste do modelo linear (por exemplo, R2). Portanto, o modelo interpretável fornece uma boa aproximação das previsões localmente.

Formalmente, uma explicação usando modelos sub-rogados locais com LIME pode ser definida como:

$$Explanation(X) = \arg\min_{g \in G} L(f, g, \pi_X) + \Omega(g)$$

onde:

f é um modelo *black box* (por exemplo, uma *random forest*), g é o modelo que explica f (por exemplo, uma regressão linear).

L é a função de perda a ser minimizada no modelo (por exemplo, erro quadrático médio), que o LIME minimiza.

Ω é a complexidade do modelo (por exemplo, número de variáveis selecionadas) decidida pelo usuário.

G é o conjunto de possíveis explicações do modelo f.

$\arg\min$ representa o valor $g \in G$ que minimiza a função $L(f,g,\pi_X) + \Omega(g)$.

π_X representa a amplitude das perturbações usadas para gerar novas observações decididas pelo usuário.

3. SHAP

SHAP[60] *(SHapley Additive exPlanations)*) é um método de explicação de modelo baseado no Teorema do Valor de Shapley[61], que foi proposto em 1952 para distribuir o valor de um jogo entre os jogadores. O SHAP é usado para explicar a importância de cada variável (medida como a alteração média na previsão do modelo quando o valor da variável varia) em uma determinada previsão.

Especificamente, o SHAP usa uma combinação de linhas de base, funções de importância local e o Teorema do Valor de Shapley para calcular a importância de cada variável em uma previsão individual.

Nesse método:

- Os valores de Shapley são calculados, onde as variáveis independentes são interpretadas como jogadores que cooperam para receber o pagamento.
- Os valores de Shapley correspondem à contribuição de cada variável para a previsão do modelo.
- O pagamento é a previsão real feita pelo modelo menos o valor médio de todas as previsões.
- Os jogadores "dividem" esse pagamento de acordo com sua contribuição, e essa divisão é calculada pelos valores de Shapley e reflete a importância de cada variável.

Esse método também permite interpretações globais, calculando a média das contribuições de cada variável para cada previsão de modelo.

Formalmente, os valores de Shapley podem ser definidos como a contribuição de cada variável para o resultado do modelo, ponderada em relação a todas as combinações possíveis de variáveis usadas:

$$\phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\}/\{j\}} \frac{|S|!(p-|S|-1)!}{p!} (val(S \cup \{j\}) - val(S))$$

em que val corresponde à previsão do modelo para variáveis incluídas no conjunto S, com relação à previsão para variáveis não incluídas em S:

$$val = \int f(x_1 \ldots x_p) dP_{x \notin S} - E_X(f(X))$$

onde:

x é o vetor de variáveis usadas no modelo.

S é um subconjunto de x.

p é o número de variáveis usadas no modelo.

$dP_{(x \notin S)}$ representa o conjunto de variáveis não incluídas em S para as quais a integração é realizada.

E eé o valor esperado da previsão de X com o modelo f.

Usando esses valores, o SHAP pode ser usado para obter uma explicação local para o modelo como:

$$Expl(x) = E_X(f(X)) + \sum \phi_j x_j$$

Por fim, o SHAP também é capaz de calcular explicações locais por meio da agregação de valores de Shapley em um conjunto de dados.

4. Anchor

O Anchors[62] é um método que explica as previsões individuais (ou seja, locais) de modelos de classificação black box encontrando regras de decisão chamadas "anchors" que explicam o resultado.

- Como no LIME, uma única previsão e os dados de entrada que a geraram são tomados como ponto de partida, e novos dados de entrada são gerados pela perturbação dessa observação, obtendo as previsões correspondentes pelo modelo AI.

[60] Lundberg, S. M., et al. (2017).). Pesquisador da Escola de Informática Paul G. Allen, Universidade de Washington.
[61] Shapley, L. (1953). Professor da Universidade da Califórnia, Los Angeles, nos Departamentos de Matemática e Economia.
[62] Ribeiro, M. T.; Singh, S.; Guestrin, C. (2018). Pesquisador da Microsoft Research no grupo de Sistemas Adaptativos e Interação e Professor Adjunto da Universidade de Washington.

- A explicação local da previsão é obtida pela busca de regras *if-else* capazes de explicar o resultado do modelo. Considera-se que uma regra explica a previsão se as alterações em outras variáveis independentes não consideradas na regra não a modificarem.

Formalmente, uma anchor A é definida como:

$$\text{Prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}[1_{f(x)=f(z)}] \geq \tau, \qquad A(x) = 1$$

onde:

f é um modelo *black box*.

D é uma distribuição arbitrária segundo a qual um distúrbio é x.

x é uma observação do conjunto de dados a ser explicado, e z é uma amostra de D.

Prec é a precisão da explicação e τ é a precisão necessária.

Uma maneira de encontrar uma âncora em uma determinada distribuição D é procurar que a precisão exceda um limite com uma certa probabilidade $(1 - \delta)$, de manera que:

$$P(Prec(A) \geq \tau) \geq 1 - \delta$$

Desenvolvimento de modelos inerentemente interpretáveis *(white box)*

Os modelos intrinsecamente interpretáveis (white box) baseiam-se no desenho de algoritmos que, por desenho, são interpretáveis e permitem que os resultados sejam explicados global e localmente.

Os modelos white box geralmente são agrupados de acordo com o tipo de algoritmo usado:

- Modelos lineares, como regressões lineares ou logísticas.
- Modelos baseados em árvores, como árvores de decisão ou árvores aleatórias.
- Modelos baseados em regras, como sistemas baseados em regras *(rule-based systems)*.
- As redes neurais profundas, com funções de ativação como ReLU ou o uso de camadas intermediárias, estão sujeitas a certas restrições que as tornam inerentemente interpretáveis[63].

[63]Yang, Z., et al. (2019). Pesquisador do Departamento de Estatística e Ciências Atuariais da Universidade de Hong Kong.

Caso de uso: Concessão de empréstimos no setor bancário. Usando SHAP].

Se o SHAP for aplicado no mesmo caso da criação de PDPs, serão obtidas informações locais adicionais sobre uma decisão no modelo para um determinado cliente.

Nesse caso, o uso do SHAP em uma amostra de observações resulta em valores de Shapley completamente diferentes com um sinal variável, dependendo das características do mutuário. Mesmo para clientes que recebem a mesma taxa de juros, a influência dessa variável varia devido à maior ou menor importância das outras variáveis no modelo.

Entretanto, observa-se uma tendência de senso comercial: quanto mais alta a taxa de juros, maior a contribuição dessa variável no modelo para uma maior probabilidade de inadimplência. Portanto, a média dos valores de Shapley de cada variável usada como uma interpretação geral do modelo pode levar a erros na explicação se for interpretada como uma generalização (Fig. 8).

Os valores de Shapley fornecem uma explicação para casos específicos, como o seguinte, em que se observa que a probabilidade de inadimplência de um cliente[1] é determinada pelas condições de hipoteca solicitadas, pelo histórico de crédito e pelas condições de emprego (por exemplo, salário) (Fig. 9).

Figura 8. Valores de Shapley para a variável "taxa de juros" em toda a amostra em relação a essa variável. O gráfico de barras cinza mostra a distribuição da variável.

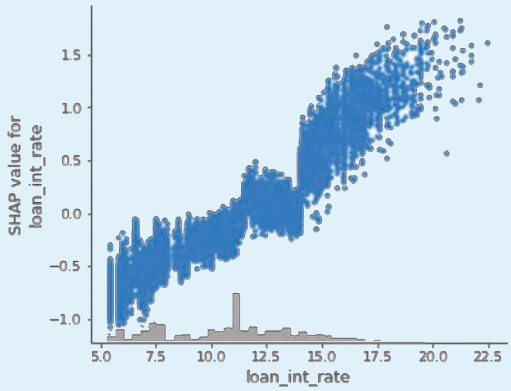
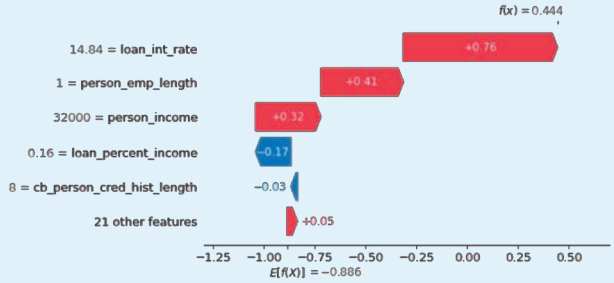
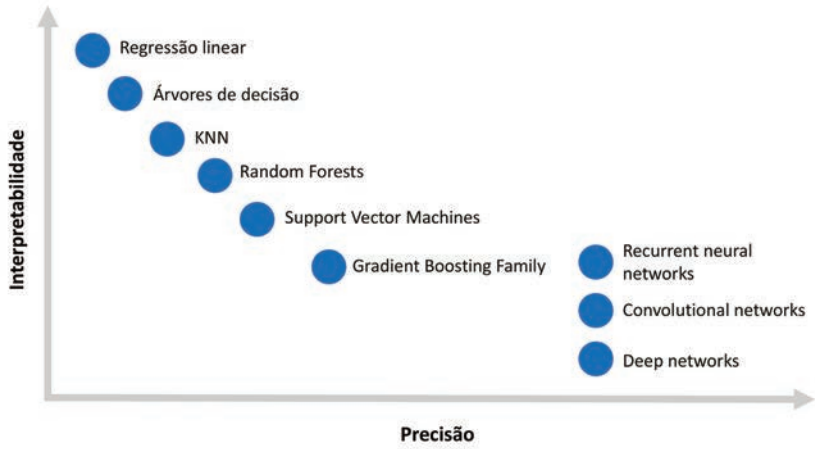


Figura 9. Valores de Shapley que influenciam a previsão de um cliente com um empréstimo negado[2].



[1]Escala do gráfico mostrada em probabilidades logarítmicas (0 corresponde a uma probabilidade de 50%).
[2]Gráfico em escala de log-odds.

Figura 10. Equilíbrio entre interpretabilidade e capacidade de previsão por famílias de modelos (incluindo white e black boxes).



O desenvolvimento desses modelos geralmente se baseia em restrições sobre os parâmetros a serem otimizados, o que permite que o modelo seja interpretável, ao contrário dos modelos black box, embora sejam menos precisos (Fig. 10). Essas restrições incluem o uso apenas de variáveis significativas para o negócio ou a restrição:

- O número de variáveis selecionadas pelo modelo para previsão.
- O número de variáveis explicadas pelo modelo.
- O grau de complexidade das regras de decisão.
- O número de etapas na previsão.
- A profundidade das árvores de decisão.
- O comprimento e a profundidade das redes neurais.

Por meio do desenvolvimento de modelos inerentemente interpretáveis, é possível obter resultados mais precisos, pois eles permitem uma melhor compreensão das informações, o que, por sua vez, possibilita uma melhor tomada de decisão. Isso é especialmente necessário nos setores em que a interpretabilidade é um fator crítico para as decisões finais.

Dois aspectos relevantes para a construção de modelos inerentemente interpretáveis são detalhados a seguir: o conceito e o desenvolvimento do aprendizado supervisionado e não supervisionado interpretável e a aplicação de outros fatores no domínio da interpretabilidade.

1. Aprendizado supervisionado e não supervisionado interpretável

Embora a pesquisa atual esteja caminhando para o desenvolvimento de modelos inerentemente interpretáveis, não existe um formalismo matemático que descreva totalmente a construção desses modelos sob quaisquer condições iniciais e algoritmos empregados.

O estado da arte é a construção desses modelos sob condições iniciais que os tornam mais facilmente interpretáveis ou equivalentes a outros modelos interpretáveis. Uma das maneiras de definir essa condição de interpretabilidade no treinamento do modelo é modificar a função de perda[64] para minimizar durante o treinamento, incluindo uma penalidade para baixa interpretabilidade, que depende de uma condição de interpretabilidade imposta no modelo f:

$$Min\left(\frac{1}{n}\sum Loss(f, z_i) + C \cdot InterpretabilityPenalty(f)\right)$$

Por exemplo, a sparsity é uma das condições usadas no desenvolvimento de modelos para qualificar um modelo como mais explicável em relação aos demais. Essa condição pode ser adicionada à função de perda como:

$$Min\left(\frac{1}{n}\sum Loss(f, z_i) + \varphi(f)\right)$$

de modo que $\varphi(f)$ é uma função de regularização que penaliza a perda por ser proporcional à esparsidade do modelo (por exemplo, se a sparsity for reduzida, esse termo da função de perda também será reduzido).

[64] Rudin, C., et al. (2022). Professor de Ciência da Computação, ECE, Estatística e Bioestatística e Bioinformática na Duke University.

Alguns autores[65] formalizaram a criação de modelos inerentemente interpretáveis para determinadas famílias, como: modelos baseados em árvores de decisão (por exemplo, SIMTree ou single-index model tree, que gera um modelo de árvore de índice único para cada nó terminal) ou a simplificação de redes com a função de ativação ReLu, que se mostra equivalente a um conjunto de modelos lineares locais.

Outros autores[66] se concentraram em definir as características que os modelos inerentemente interpretáveis devem ter para otimizá-los durante o processo de modelagem, como, por exemplo:

- Aditividade das variáveis de entrada, de modo que seus efeitos sejam agregados no modelo de forma simples.
- Sparsity e a otimização de modelos para atender a essa condição.
- Linearidade das variáveis de entrada versus saída do modelo.
- Monotonicidade, de modo que, para o maior número possível de intervalos, a relação entre a variável de entrada e o resultado a ser previsto seja monotônica.
- Desacoplamento de conceitos no treinamento de redes neurais, que se refere a manter, tanto quanto possível, as informações sobre um determinado conceito em caminhos específicos na rede (ou seja, em face das informações sobre o mesmo conceito que atravessam um número maior de neurônios e caminhos dispersos na rede).
- Redução de dimensionalidade como uma ferramenta visual para facilitar explicações post-hoc para humanos.

2. Outros fatores de impacto

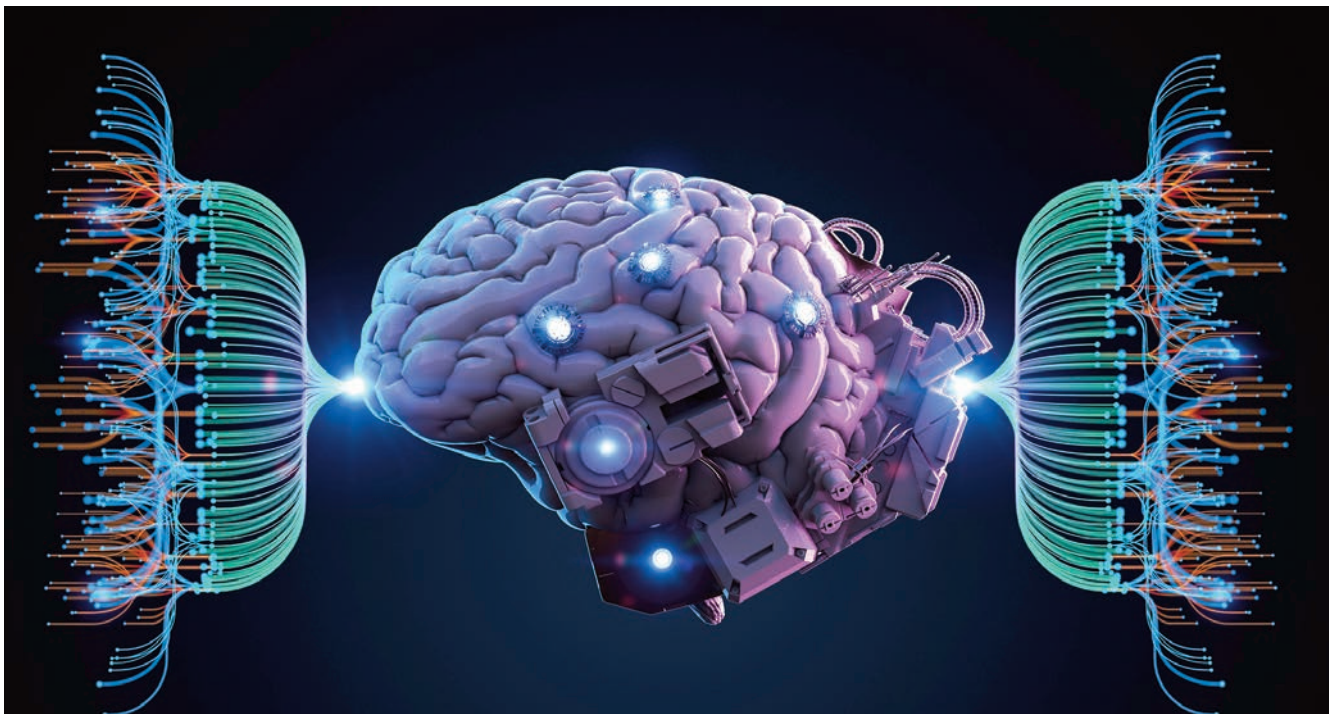
Em combinação com os desafios mostrados nesta seção, há outros elementos-chave que podem ser considerados para melhorar a interpretabilidade do modelo, como a imparcialidade do modelo, a ausência de viés nos dados de entrada, componentes especializados em potencial ou desempenho adequado e estrutura de controle do modelo para evitar erros na interpretação do modelo.

Devido à sua relevância, conforme indicado acima[67], esses elementos também foram destacados no AI Act como requisitos essenciais para sistemas de AI de alto risco.

Atualmente, existem várias técnicas e métodos para avaliar o desempenho dos modelos e evitar problemas de overfitting. Há também várias maneiras de avaliar o erro produzido pelo modelo e o equilíbrio entre o viés e o erro de variância. No entanto, devido às limitações no uso de dados pessoais introduzidas pelas normas de proteção de dados, uma das maiores complexidades no momento é detectar e corrigir possíveis vieses (por exemplo, por raça, gênero, religião, orientação política ou sexual, crenças ou posição social) nos modelos de AI, especialmente quando as variáveis não são armazenadas e, portanto, não estão disponíveis para análise.

Nesse sentido, várias técnicas para identificar variáveis de entrada imparciais foram propostas no meio acadêmico, como:

[65] Sudjianto. A., et al. (2021).
[66] Rudin, C., et al. (2022).
[67] Ver a seção sobre regulamentação.



- Análise de interpretabilidade por meio de redes causais bayesianas[68] como uma quantificação do grau de imparcialidade do modelo.

- Definição[69] de métricas de imparcialidade, como paridade demográfica, paridade de proporção preditiva, falsos positivos e falsos negativos iguais em segmentos suscetíveis a viés.

Entre essas métricas está a imparcialidade contrafactual *(counterfactual fairness),* que fornece uma medida da semelhança dos resultados de um modelo com indivíduos (observações) com as mesmas características, mas com atributos sensíveis a vieses ligeiramente diferentes.

Vantagens e desvantagens das técnicas de interpretabilidade mais comuns

Como regra geral, não existe uma técnica de interpretabilidade que possa fornecer uma explicação única, abrangente e intuitiva para qualquer cenário. As técnicas de interpretabilidade são frequentemente combinadas em vários casos de uso e cenários para verificar se fornecem explicações reproduzíveis aplicáveis a diferentes conjuntos de observações.

Ao selecionar qual dessas técnicas usar, é útil considerar as vantagens ou desvantagens de sua aplicação (Fig. 11).

Últimas tendências e desafios

Apesar dos avanços na interpretabilidade do modelo, ainda há desafios para explicar os resultados (Fig. 12).

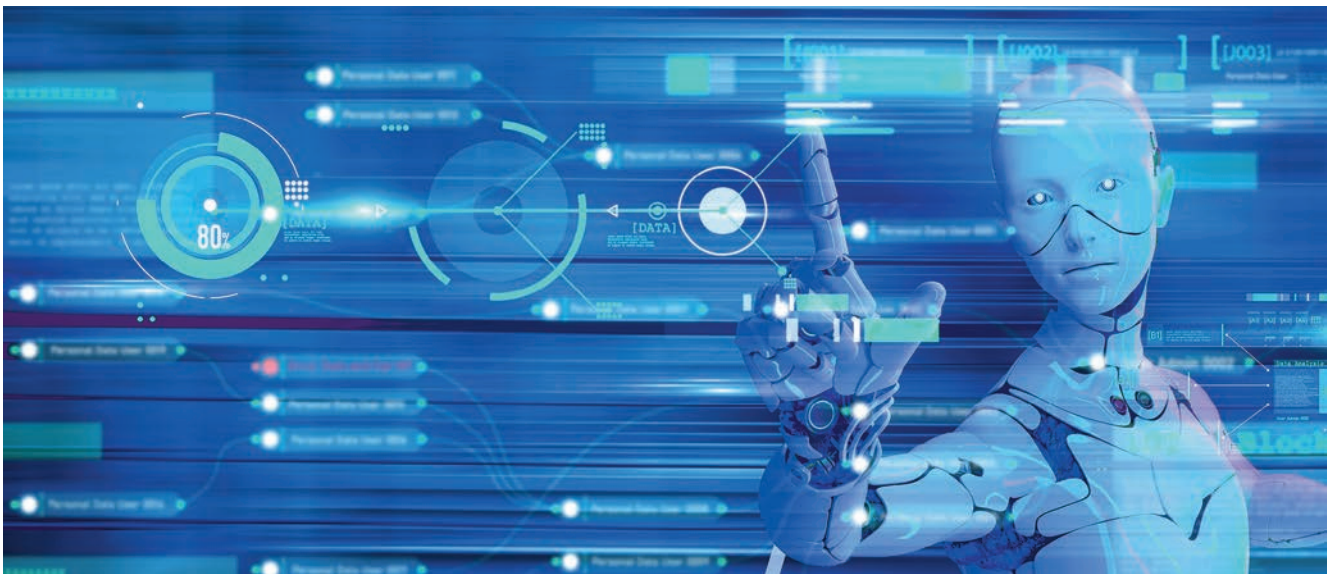
Em primeiro lugar, a interpretabilidade dos modelos ainda é limitada por vários fatores, como a reprodutibilidade dos resultados[70], o processo de treinamento e implementação do modelo, a consistência de suas previsões, a explicação da sequência de previsões mais prováveis, os vieses nos dados de entrada, a imparcialidade (fairness) e a precisão da explicação.

Em segundo lugar, as técnicas de XAI atualmente disponíveis permitem apenas explicações locais (ou seja, para uma única observação ou dado) ou globais (ou seja, para todo o conjunto de dados). Isso cria a necessidade de desenvolver técnicas que permitam explicações em configurações intermediárias, ou seja, explicar resultados para grupos ou subconjuntos de dados[71]. Além disso, sem uma análise aprofundada, os resultados de diferentes técnicas de interpretabilidade em diferentes níveis podem inicialmente parecer contraditórios (por exemplo, se alguém tentar comparar resultados globais "médios" com resultados locais em uma configuração).

[68] Oneto, L.,Chiappa, S., (2020)
[69] Zhou, N., et al. (2021). Analista financeiro sênior da Wells Fargo.
[70] Leventi-Peetz, A.-M., et al. (2022).
[71] Embora o SHAP consiga obter explicações para subconjuntos por meio de médias ponderadas dos valores de Shapley, é possível que essas explicações variem dependendo da granularidade dos dados do subconjunto.

Figura 11. Comparação das técnicas de interpretabilidade mais comuns.

Técnica	Vantagens	Desvantagens
1 PDP (Partial Dependence Plot)	✓ Fácil de aplicar e intuitivo de implementar. ✓ O cálculo dos gráficos de dependência parcial tem uma interpretação causal.	X Por definição, ele não permite que o impacto de mais de duas variáveis seja visto intuitivamente no gráfico. X Ele não explica como a explicação varia de acordo com uma única variável independente se todas as outras variáveis independentes variarem.
2 LIME (Local interpretable model-agnostic explanations)	✓ Com base em uma previsão, esse método avalia o impacto de pequenas alterações nos insumos. ✓ Um modelo substituto local é usado para avaliar as diferenças entre as previsões originais e modificadas, bem como as variáveis mais importantes que contribuem para a previsão. ✓ O método é agnóstico com relação ao modelo de previsão usado. ✓ Pressupõe-se linearidade local.	X Pressupõe-se linearidade local. X Ele pode gerar explicações contrárias em diferentes subconjuntos de dados, portanto, é necessário verificar as explicações em intervalos representativos do conjunto de dados. X Ele não fornece uma explicação geral do modelo.
3 SHAP (SHapley Additive exPlanations)	✓ Calcula a contribuição de cada variável para uma previsão específica. ✓ Ele não pressupõe linearidade local. ✓ Ele pode abranger a importância geral das características para todo o conjunto de dados. ✓ Agnóstico com relação ao modelo de previsão usado. ✓ É muito caro do ponto de vista computacional e pressupõe que as variáveis do modelo sejam independentes.	X Ele pode gerar explicações contrárias em diferentes subconjuntos de dados, portanto, é necessário verificar as explicações em intervalos representativos do conjunto de dados. X Ele não fornece uma explicação geral do modelo.
4 Anchors	✓ Não depende do tipo de modelo e é fácil de interpretar. ✓ Ele captura o comportamento não linear de modelos complexos.	X Grande número de hiperparâmetros (forma de perturbação, precisão...) X Isso requer a discretização de variáveis contínuas em muitos casos, o que pode levar a erros de interpretação.
5 *Construção de modelos "white box"*	✓ Reduz o esforço de interpretação do modelo após o treinamento e durante seu ciclo de vida. ✓ Isso não leva a contradições na interpretação do modelo e facilita seu uso. ✓ Ele não requer o uso de modelos ou técnicas post-hoc adicionais.	X Aumenta o esforço durante a construção do modelo. X Não há técnicas aplicáveis a todos os tipos de modelos nesta fase.



Em terceiro lugar, ainda são necessários aprimoramentos no desenvolvimento de modelos white box, pois, apesar do progresso feito nos últimos anos, esses modelos ainda não são capazes de competir em precisão com os modelos black box em problemas complexos.

Por fim, a necessidade de explicar modelos mais complexos (por exemplo, determinados tipos de redes neurais profundas) continua sendo um desafio não resolvido.

Nesse sentido, novas técnicas estão sendo desenvolvidas para melhorar a interpretabilidade dos modelos, como o uso de informações das camadas intermediárias de redes neurais profundas, a agregação de métricas de interpretabilidade para medir a explicabilidade dos modelos, o desenvolvimento de modelos adversários para quantificar o grau de explicabilidade, a limitação dos parâmetros a serem otimizados para aumentar sua interpretabilidade ou o uso de técnicas de visualização para facilitar a compreensão dos resultados.

Figura 12. Desafios comuns na interpretabilidade dos modelos de AI.

	Desafios	Soluções
Interpretabilidade do modelo	**O processo de modelagem deve ser reproduzível,** apesar da variação inerente dos modelos de AI.	Controle de randomização em variáveis e algoritmos (por exemplo, uso de sementes).
	As explicações de interpretabilidade **podem não ser exclusivas, intuitivas,** explicar totalmente a previsão ou ser consistentes com dados semelhantes.	Comparação de várias técnicas de interpretabilidade aplicadas em amostras representativas, verificando a consistência em dados semelhantes. Criação de relatórios de interpretabilidade compreensíveis por áreas menos técnicas.
	A sequência de previsões de modelos mais prováveis pode não ser consistente, e as previsões podem ser sensíveis a **perturbações nos dados de entrada**.	Aplicação de técnicas de interpretabilidade na sequência mais provável de previsões do modelo e inclusão de análises de sensibilidade específicas ao revisar o modelo.
	Não existe um **formalismo matemático para descrever as propriedades de muitos modelos de ML** (por exemplo, em modelos de aprendizagem profunda).	Documentação abrangente sobre os modelos e suas bases matemáticas, além de refletir o aumento do risco na classificação por níveis.
Risco de modelo	Não há **medidas de avaliação globalmente aceitas** para modelos de ML, nem de interpretabilidade.	Aplicação de uma estrutura de testes diversificada e abrangente em todo o ciclo de vida do modelo.
	Os usuários do modelo devem ser **capazes de acessar os módulos de interpretabilidade para entender as previsões,** mesmo após o desenvolvimento.	Implementação de módulos de XAI que dão acesso aos resultados de interpretabilidade para diferentes usuários.
Dados, documentação e implementação	Há uma **variação de implementação em determinados algoritmos de AI.**	Documentação adequada e controles de implementação.
	O nível de **documentação deve refletir a complexidade do algoritmo.**	Inclua em detalhes os resultados dos testes de interpretabilidade, garantindo sua replicação por terceiros.