ManagementSolutions
Making things happen

Explainable artificial intelligence (XAI)
Desafios na interpretabilidade de modelos

P&D

www.managementsolutions.com

Design e diagramação
Departamento de Marketing e Comunicação
Management Solutions - Espanha

Fotograflas

Arquivo fotográfico da Management Solutions
iStock

Adobe stock

Índice

Introdução

"A maior parte do que fazemos com o machine learning acontece abaixo da superfície. Embora possa não ser visível, grande parte do impacto do machine learning será assim: uma melhoria silenciosa, mas significativa, das operações essenciais."

Jeff Bezos[1]

"A inteligência artificial (AI) é o campo da ciência e da engenharia voltado para a criação de máquinas inteligentes e, principalmente, de software inteligente. Ela está relacionada à tarefa semelhante de usar computadores para entender a inteligência humana, mas a AI não precisa se limitar a métodos biologicamente observáveis"[2].

Essa foi a definição de AI oferecida por John McCarthy, professor da Universidade de Stanford, um dos fundadores da disciplina e coautor do termo "inteligência artificial".

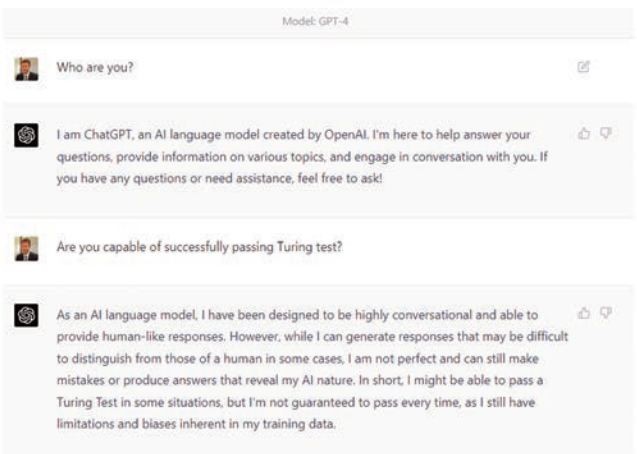
No entanto, já em 1950, Alan Turing se perguntava[3]: "as máquinas podem pensar?" e formulou o que mais tarde se tornaria conhecido como o "teste de Turing": um teste da capacidade de uma máquina de demonstrar inteligência indistinguível da de um ser humano. Turing propôs que um avaliador humano julgasse as conversas em linguagem natural entre uma pessoa e uma máquina projetada para gerar respostas semelhantes às humanas. Se o avaliador não conseguisse distinguir a máquina do ser humano, a máquina teria passado no teste.

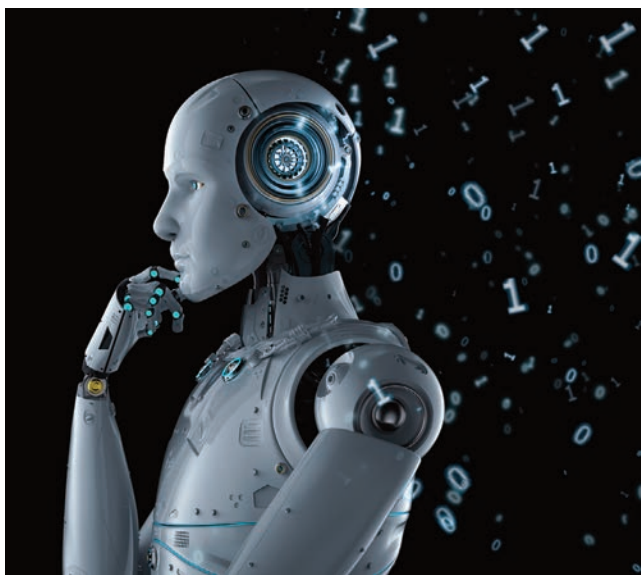
Embora haja controvérsia sobre isso[4], muitos autores consideram que já existem inteligências artificiais que poderiam passar no teste de Turing, como o GPT-4, da Open AI Foundation, embora o próprio GPT-4 não seja totalmente claro (Fig. 1). Há também testes mais sofisticados, como o desafio de esquemas de Winograd, que consiste em resolver anáforas complexas que exigem conhecimento e bom senso[5], algo que a AI atual ainda não parece ser capaz de fazer.

[1]Bezos (nascido em 1964), J., fundador, presidente executivo e ex-CEO da Amazon.
[2]McCarthy (2004). Professor de Ciência da Computação na Universidade de Stanford.
[3]Turing (1950). Matemático, lógico, cientista teórico da computação, criptógrafo, filósofo e biólogo teórico britânico.
[4]Harnad (2003). Professor de Psicologia da Universidade de Quebec em Montreal (UQAM) e da Universidade McGill, e Professor Emérito de Ciências Cognitivas da Universidade de Southampton.
[5]Um esquema Winograd é uma pergunta de escolha binária em que (i) há duas partes mencionadas na pergunta; (ii) pronomes são usados para se referir a elas; (iii) há ambiguidade sobre a quem o pronome se refere; e (iv) há palavras específicas que podem alterar a resposta correta. Em um exemplo do mesmo Terry Winograd (Professor de Ciência da Computação da Universidade de Stanford):
- Pergunta: Os conselheiros municipais negaram permissão aos manifestantes porque (temiam/defendiam) a violência. Quem (teme/defende) a violência?
- Resposta: (os conselheiros / os manifestantes).

Con ello, se puede generar un test alternativo al Test de Turing, utilizando dichas preguntas y penalizando fuertemente las respuestas erróneas (véase Levesque (2014)).

Figura 1. Conversa com o GPT-4 sobre sua capacidade de passar no teste de Turing.





Ainda assim, embora o campo da AI não seja novo, nos últimos anos houve avanços vertiginosos, com aplicações que vão de carros autônomos a diagnósticos médicos, trading automático, reconhecimento facial, gerenciamento de energia, segurança cibernética, robótica e tradução automática, para citar apenas alguns.

Uma característica distintiva da AI atual está precisamente ligada à definição de McCarthy acima: ela não se limita a métodos observáveis e, quando atinge um certo nível de sofisticação, apresenta problemas de interpretabilidade. Em outras palavras: os modelos de AI tendem a ter uma alta taxa de acerto, muito maior do que os algoritmos tradicionais; mas, em cada caso individual, pode ser extremamente complexo explicar por que o modelo produziu um determinado resultado.

Embora existam aplicações de AI em que seja menos relevante poder entender ou explicar por que o algoritmo retornou um determinado valor, em muitos casos isso é essencial e é um requisito regulatório. Por exemplo, na União Europeia, de acordo com o Regulamento Geral de Proteção de Dados (GDPR), os consumidores têm o que é conhecido como "direito a uma explicação"[6]:

> [...] não estar sujeito a uma decisão [...] baseada exclusivamente em processamento automatizado [...], como a rejeição automática de uma solicitação de crédito on-line, [...] [na qual] não há intervenção humana envolvida", e tem o direito de "receber uma explicação sobre a decisão tomada [...] e contestar a decisão".

Isso levou ao desenvolvimento da disciplina de Inteligência Artificial Explicável (XAI), que é o campo de estudo que visa tornar os sistemas de AI compreensíveis para os seres humanos[7], em oposição à noção de "black box", que se refere a algoritmos nos quais somente os resultados são observáveis e o funcionamento do modelo é desconhecido, ou a lógica dos resultados não é explicada.

Um algoritmo pode ser considerado[8] como pertencente à disciplina XAI se seguir três princípios: transparência, interpretabilidade e explicabilidade. A transparência é garantida se os processos que calculam os parâmetros do modelo e produzem os resultados puderem ser descritos e justificados. A interpretabilidade descreve a possibilidade de entender o modelo e apresentar como ele toma decisões de uma forma compreensível para o ser humano. A explicabilidade refere-se à capacidade de decifrar por que uma determinada observação recebeu um valor específico. Na prática, esses três termos estão intimamente ligados e são frequentemente usados de forma intercambiável, na ausência de consenso sobre suas definições precisas[9].

Para alcançar esses princípios, há basicamente duas abordagens possíveis: desenvolver algoritmos que sejam interpretáveis e explicáveis por sua natureza (como regressões lineares, modelos logísticos ou multinomiais ou certos tipos de redes neurais

[6]GDPR (2018), Considerando 71.
[7]Vilone et al. (2021). Doutorado em Inteligência Artificial, School of Computer Science, Technological University of Dublin.
[8]Roscher et al. (2020). Cientista de dados da Universidade Técnica de Munique.
[9]Marcinkevics et al. (2020). Pesquisador do Departamento de Ciência da Computação, ETH Zurich.

profundas, entre outros) ou usar técnicas de interpretabilidade como ferramentas para alcançar esses princípios[10].

Portanto, a XAI se preocupa tanto com técnicas para tentar explicar o comportamento de determinados modelos opacos ("black box") quanto com o design de algoritmos inerentemente interpretáveis ("white box")[11].

A XAI é fundamental para o desenvolvimento da AI e, portanto, para os profissionais que trabalham em contato com ela, devido a pelo menos três fatores:

- Contribui para gerar confiança na tomada de decisões com base em modelos de AI; sem essa confiança, os usuários desses modelos podem demonstrar resistência à sua adoção.
- É um requisito regulatório em determinadas áreas (por exemplo, proteção de dados, proteção ao consumidor, igualdade de oportunidades na contratação de funcionários, regulação de modelos em bancos).
- Favorece o aprimoramento e a robustez dos modelos de AI (por exemplo, identificando e eliminando vieses, compreendendo as informações relevantes para produzir um determinado resultado ou antecipando possíveis erros em observações não contempladas na amostra de treinamento do modelo). Isso leva ao desenvolvimento de algoritmos éticos e permite que as organizações concentrem seus esforços na identificação e na garantia da qualidade dos dados que são relevantes para seus processos de tomada de decisão.

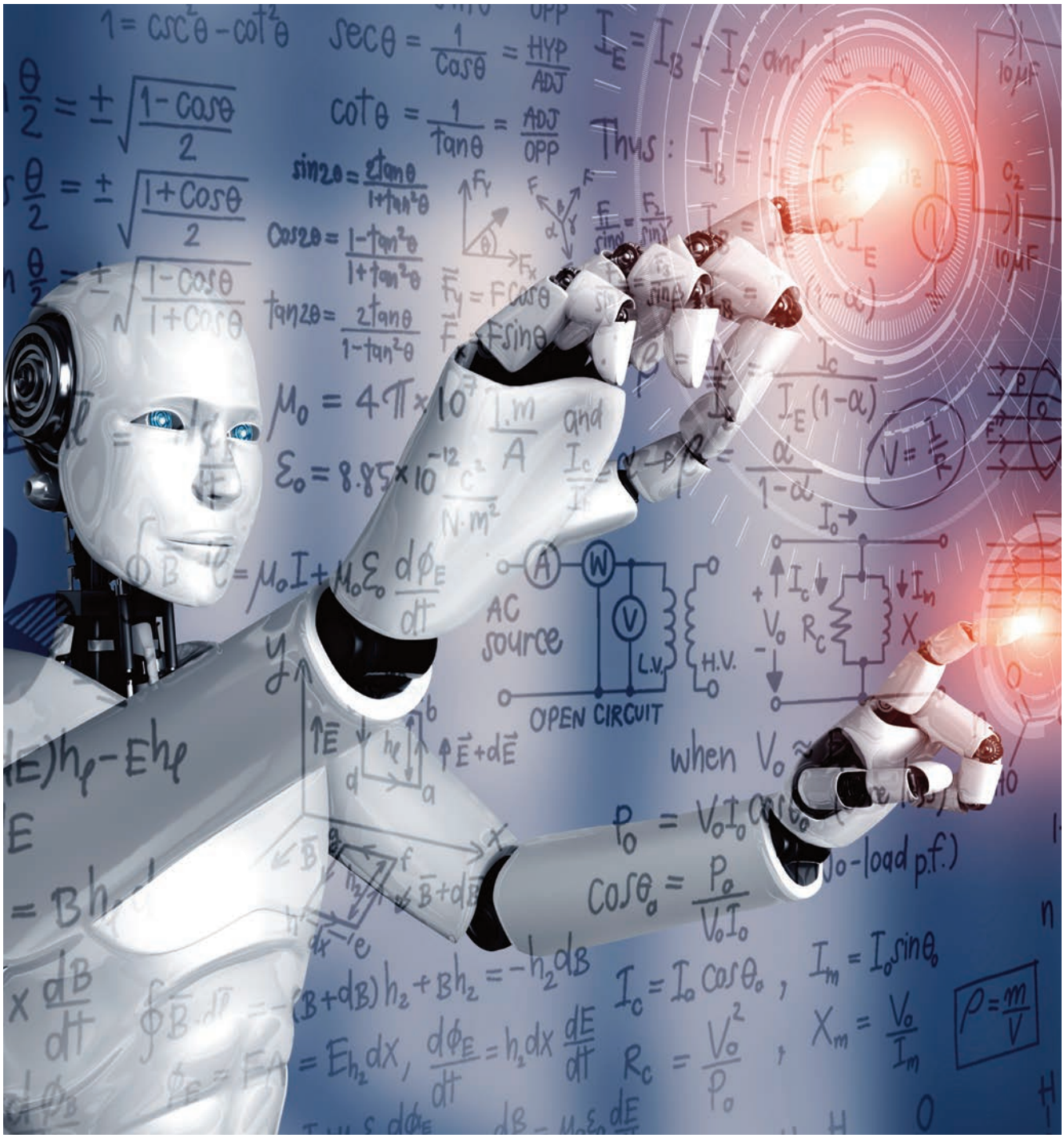
Embora o desenvolvimento de sistemas XAI esteja recebendo muita atenção do meio acadêmico, do setor e dos órgãos reguladores, ele ainda apresenta vários desafios.

Este documento analisará o contexto e a justificativa da XAI, incluindo a regulação sobre o assunto e suas implicações na organização; o estado da arte e as principais técnicas da XAI; e o progresso e os desafios não resolvidos da XAI. Por fim, será apresentado um estudo de caso da XAI para ajudar a ilustrar sua aplicação prática.

[10] iDanae (2022). Cátedra (inteligência, dados, análise e estratégia) em Big Data e Analytics, que surgiu graças à colaboração entre a Management Solutions e a Universidade Politécnica de Madri (UPM) nos âmbitos educativo, científico e técnico, e tem como objetivo promover a geração de conhecimento, a difusão e a transferência de tecnologia, e fomentar a P&D&I na área de Data Analytics.
[11] Sudjianto et al. (2011). Diretor de Risco de Modelo da Wells Fargo.

Resumo executivo

"Toda tecnologia deve ser acompanhada de um manual especial: não sobre como usá-la, mas por quê, quando e para quê."
Alan Kay[12]



Contexto e fundamentos da XAI

1. A transformação digital possibilitou o acesso e a exploração de uma grande quantidade de dados estruturados e não estruturados, impulsionando o uso de técnicas de machine learning e inteligência artificial em todos os setores.

2. Os modelos de AI oferecem maior poder de previsão, mas também apresentam riscos, como a presença de vieses inadvertidos, a falta de compreensão do modelo ou erros em sua aplicação decorrentes de causas como o treinamento excessivo, o que pode levar à desconfiança em relação ao modelo. Isso levanta a questão de saber se é possível entender os resultados dos algoritmos de AI suficientemente bem para tomar as decisões adequadas.

3. A Inteligência Artificial Explicável (XAI) é um conjunto de processos e métodos que permite que os usuários entendam e confiem nos resultados e saídas criados pelos algoritmos de machine learning. Essa disciplina é fundamental para que uma organização crie confiança ao empregar modelos de AI, ajudando a caracterizar a precisão do modelo, a imparcialidade, a transparência e a compreensão dos resultados na tomada de decisões baseada em AI.

4. O interesse acadêmico e profissional na XAI aumentou exponencialmente nos últimos anos, devido à capacidade da disciplina de abordar uma série de preocupações do setor no uso da AI, como requisitos regulatórios, falta de confiança, potencial uso indevido, impacto reputacional, impactos sociais ou humanos e outros riscos.

5. Isso levou os órgãos reguladores e supervisores de diferentes jurisdições a estabelecer regulamentos e diretrizes para o uso adequado da AI, incluindo aspectos de interpretabilidade do modelo.

6. Na Europa, o Regulamento Geral sobre a Proteção de Dados (GDPR) do Parlamento Europeu, que entrou em vigor em 2018, estabelece um "direito a uma explicação" para os cidadãos, exigindo que as empresas possam explicar por que um modelo de AI retornou um determinado resultado. Isso tem implicações críticas para o design e a análise de interpretabilidade dos modelos de AI.

7. Além disso, o Parlamento Europeu propôs em 2021 o Artificial Intelligence Act (AI Act) para regulamentar o uso da inteligência artificial na União Europeia. Esse regulamento proposto estabelece um framework regulatório para sistemas de AI, incluindo requisitos de desenvolvimento ético, transparência, segurança e precisão, bem como um sistema de governança e supervisão. A Lei de AI classifica os aplicativos de AI em níveis de risco (práticas inaceitáveis, sistemas de alto risco e sistemas de risco baixo ou limitado) e estabelece requisitos de transparência e supervisão humana para sistemas de alto risco, que serão aplicáveis em toda a União. É provável que isso desencadeie iniciativas de adaptação ao Regulamento, como documentação abrangente de modelos, técnicas de interpretabilidade, dashboards de monitoramento e alertas de modelos, entre outros.

8. Além disso, a Comissão Europeia formulou em 2019 as Diretrizes Éticas para Inteligência Artificial Confiável, que propõem sete requisitos principais para que os sistemas de AI sejam considerados confiáveis: (i) ação e supervisão humanas, (ii) robustez técnica e segurança, (iii) gestão da privacidade e dos dados, (iv) transparência, (v) diversidade, não discriminação e equidade, (vi) bem-estar ambiental e social e (vii) responsabilização. Dentro do requisito de transparência, é estabelecida a necessidade de explicabilidade dos modelos de AI. As Diretrizes propõem critérios de avaliação para determinar até que ponto um modelo de AI atende a esses requisitos.

9. Nos Estados Unidos, em 2022, a Casa Branca propôs uma minuta da Declaração de Direitos sobre a AI (AI Bill of Rights), promovida pelo presidente Joe Biden. Esse projeto de lei estabelece cinco princípios ou direitos dos cidadãos com relação à AI, incluindo sistemas seguros e eficazes, proteção contra discriminação dos algoritmos, privacidade de dados, notificação e explicação, e avaliação e correção humanas em caso de falha da AI (fallback). Esses princípios incluem a explicabilidade dos modelos de AI, que exige

[12] Alan Kay (nascido em 1940), cientista da computação americano ganhador do Prêmio Turing, considerado o "pai dos computadores pessoais".

documentação em linguagem simples, explicações tecnicamente válidas, significativas e úteis, e notificações de uso comprovadamente claras, oportunas, compreensíveis e acessíveis.

10. Os Princípios da OCDE sobre Inteligência Artificial de 2019 promovem o uso de AI que seja confiável e respeite os direitos humanos e os valores democráticos. Eles foram adotados por todos os 38 países membros da OCDE e exigem, entre outros, transparência e divulgação responsável dos sistemas de AI para que as pessoas afetadas por um sistema de AI possam compreender o resultado.

11. O Discussion Paper on Machine Learning for IRB Models[da Autoridade Bancária Europeia (EBA), publicado em 2021, analisa a relevância de possíveis barreiras à implementação de técnicas de machine learning na abordagem IRB para o cálculo de capital em instituições financeiras. O documento estabelece princípios e recomendações para tornar o uso dessas técnicas compatível com a conformidade com a regulação europeia de requisitos de capital (CRR). Essas recomendações incluem análise estatística e econômica da relação entre as variáveis de entrada e a variável de saída, documentação que explique o modelo de forma simples e a necessidade de detecção de possíveis vieses no modelo.

12. Um princípio básico da XAI é a necessidade de integrar a interpretabilidade e a explicabilidade à organização e aos processos de uma empresa. Isso é feito por meio de um framework de XAI que consiste em quatro elementos: técnicas de interpretabilidade de modelos de AI, integração em processos de gestão de risco de modelo (MRM), suporte tecnológico e fatores humanos.

13. Técnicas: o núcleo do framework de XAI baseia-se em três aspectos principais de interpretabilidade: a explicação do desenho do modelo, a explicação dos resultados do modelo e outros aspectos, como a detecção de viés e o monitoramento periódico do modelo.

14. MRM: A interpretabilidade dos modelos de AI é uma característica que afeta toda a cadeia do ciclo de vida do modelo e, portanto, a gestão do risco do modelo. Para incorporar elementos XAI, a estrutura organizacional e de governança, as políticas e os procedimentos para desenvolvimento, monitoramento, validação, implementação e uso de modelos, bem como a estrutura de auditoria, precisam ser revisados e atualizados.

15. Suporte tecnológico: para implementar um framework de XAI, são necessárias soluções tecnológicas profissionais para dar suporte aos aspectos de interpretabilidade dos modelos de AI, como ferramentas de interpretabilidade e de governança de modelos, sistemas de análise de dados, APIs, mecanismos de segurança e auditoria e protocolos para garantir a conformidade com os padrões de qualidade e explicabilidade.

16. Fator humano: a integração da XAI deve considerar o fator humano, incluindo o recrutamento e a retenção de talentos especializados, programas de treinamento, a criação de uma cultura que aprimore o uso da explicabilidade e interpretabilidade dos modelos de AI e programas de gestão de mudanças para garantir a adoção adequada da XAI.

17. Além disso, um quinto elemento central para a AI e a XAI são os dados, pois sua governança, qualidade, integridade, consistência, rastreabilidade e ausência de viés determinam a qualidade do modelo de AI e, em última análise, das decisões tomadas com base nele. No entanto, as questões relacionadas aos dados e sua relevância para os modelos não são o tema deste documento, pois já foram amplamente abordadas em publicações anteriores[13].

Técnicas de interpretabilidade: estado da arte

18. O uso de técnicas de AI se espalhou por todos os setores e domínios, oferecendo maior poder de previsão em troca de maior complexidade. Isso gerou a necessidade de explicar os resultados dos modelos de AI, o que levou ao surgimento de técnicas cada vez mais sofisticadas de interpretabilidade local e global. Essas técnicas não resolvem completamente o problema, e diferentes abordagens para garantir a interpretabilidade dos modelos de AI continuam a ser desenvolvidas, como o desenvolvimento de modelos inerentemente interpretáveis ("caixas brancas").

19. As abordagens mais comuns para tratar o problema da interpretabilidade podem ser classificadas em dois grupos: interpretabilidade post-hoc (técnicas de interpretabilidade global e local) e modelos inerentemente interpretáveis. Além disso, há estratégias complementares, como a simplificação do modelo, o uso de variáveis de senso comercial, a análise de dados para identificar parcialidade ou imparcialidade e a análise da reprodutibilidade do desenvolvimento do modelo.

20. O LIME (Local Interpretable Model-agnostic Explanations) permite que um modelo seja explicado de forma local e agnóstica, ou seja, ele pode gerar explicações para uma previsão específica sem a necessidade de entendimento do modelo subjacente.

21. O SHAP (SHapley Additive exPlanations) explica o modelo de forma global, avaliando a contribuição de cada variável de entrada para a previsão de saída.

22. Os PDPs (Partial Dependence Plots, gráficos de dependência parcial) são usados para visualizar como o resultado de um modelo muda quando os valores das variáveis de entrada são alterados.

[13] Ver Management Solutions (2020, 2018 e 2015): "Auto machine learning, rumo à automação de modelos", "Machine learning, uma peça-chave na transformação dos modelos de negócios" e "Data science e a transformação do setor financeiro".

23. Os modelos white box baseiam-se no desenvolvimento de algoritmos que, por sua concepção, são inerentemente interpretáveis. Esses modelos são agrupados de acordo com o tipo de algoritmo usado, e os parâmetros a serem otimizados geralmente são limitados para obter maior interpretabilidade. Isso resulta em resultados mais precisos, pois permite uma melhor compreensão das informações, o que, por sua vez, resulta em uma melhor tomada de decisão, especialmente nos setores em que a interpretabilidade é um fator crítico.

24. Apesar dos avanços na interpretabilidade dos modelos de AI, ainda há desafios, como a reprodutibilidade dos resultados, a explicação da sequência de previsões mais prováveis, vieses nos dados de entrada, imparcialidade (fairness) e precisão da explicação. Além disso, há espaço para melhorias no desenvolvimento de modelos white box para competir em precisão com modelos black box em problemas complexos, bem como no desenvolvimento de novas técnicas para explicar modelos mais complexos.

Estudo de caso de interpretabilidade

25. Para demonstrar a aplicação das técnicas de interpretabilidade descritas acima, é realizado um exercício ilustrativo usando dados fictícios gerados pela IBM e publicados no Kaggle[14]. O objetivo do estudo é entender as causas que levam os funcionários a deixar seus empregos, usando técnicas de AI e XAI nos dados fictícios propostos.

26. O exercício foi realizado com a ajuda de um sistema de modelagem de componentes, o ModelCraft™[15], que contém múltiplas técnicas relevantes de AI e XAI, permitindo que o estudo seja concluído em muito menos tempo do que o normal e sem a necessidade de escrever código.

27. Para explicar o desgaste dos funcionários, diferentes modelos foram treinados e validados, entre os quais o algoritmo de random forest demonstrou a melhor capacidade de previsão.

28. Para explicar os resultados do modelo, foram aplicadas as técnicas de interpretabilidade SHAP, LIME e PDP para entender quais variáveis explicam melhor a fuga dos funcionários, como as mudanças nas variáveis mais importantes afetam diferentes faixas populacionais e os resultados do modelo em casos individuais.

29. A aplicação e a interpretação corretas do modelo nesse estudo de caso permitiriam antecipar e evitar a rotatividade de funcionários, criar perfis com diferentes propensões à fuga e identificar antecipadamente as características desses funcionários para tomada de medidas adequadas. Além disso, esse caso de uso destaca as limitações e as dificuldades na aplicação de técnicas de interpretabilidade post-hoc, bem como o fato de que o uso de modelos de AI juntamente com um módulo de interpretabilidade pode aumentar a capacidade preditiva do modelo.

Conclusão

30. A Inteligência Artificial Explicável (XAI) é uma disciplina emergente que busca melhorar a interpretabilidade dos modelos de AI usando técnicas específicas para entender e explicar os resultados dos modelos de AI, e é especialmente importante em âmbitos de alta sensibilidade, como saúde, segurança, serviços financeiros e de energia, entre outros.

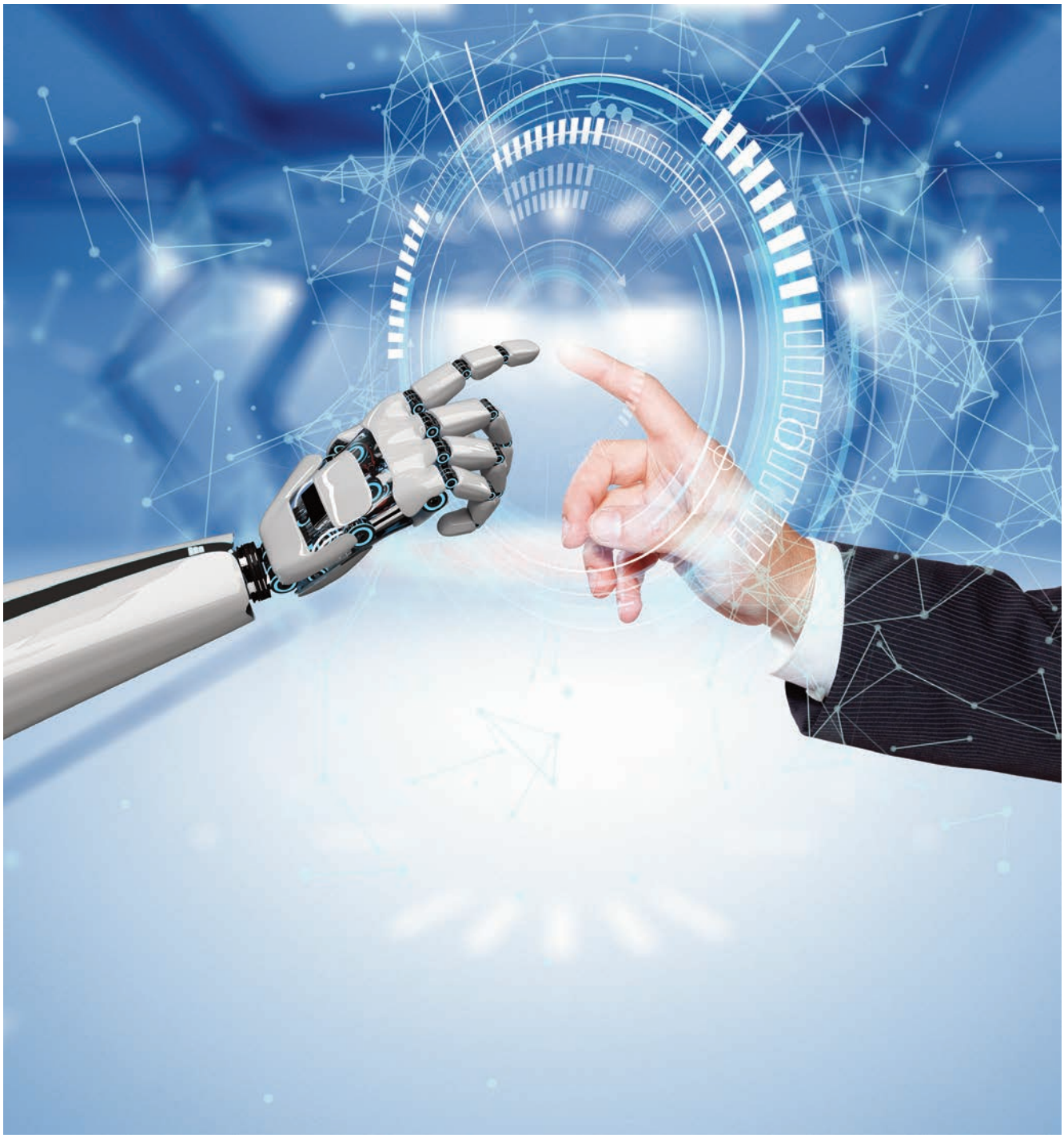
31. A XAI se tornou uma prioridade para muitos setores à medida que os modelos de AI se tornam cada vez mais complexos e cada vez mais regulações exigem sua interpretabilidade. Um estudo de caso desenvolvido com o ModelCraft™ demonstrou como essas técnicas podem ser usadas para entender e explicar os modelos de AI.

32. Nos próximos anos, espera-se que a XAI continue a se desenvolver e a crescer em importância à medida que os modelos de AI se tornam mais complexos, a regulação continue a proliferar e seu uso se estenda para mais áreas de alta sensibilidade.

[14] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.
[15] Ferramenta de AutoML e de modelagem por componentes proprietária da Management Solutions. Ver Management Solutions (2023).

Contexto e fundamentos da XAI

"Compreender a inteligência artificial é um desafio que exige uma enorme capacidade intelectual; felizmente, temos a inteligência artificial para lidar com isso."
GPT-4[16]



Contexto

Um dos recursos mais notáveis da transformação digital é que ela está disponibilizando para todos os setores uma quantidade enorme de dados estruturados e não estruturados provenientes de vários aplicativos; por exemplo:

- Dados de varejo procedentes de ações de compra, transações e feedbacks dos clientes.
- Dados financeiros de fontes bancárias, de investimento e comerciais.
- Dados de redes sociais, incluindo análise de opiniões e análise preditiva.
- Sensores digitais de IoT (Internet das Coisas) que medem temperatura, pressão e outros dados do entorno.
- Dados de saúde, como registros médicos, diagnósticos, imagens e informações genômicas.
- Wearables, como rastreadores de atividade, sensores de saúde e smartwatches.
- Sistemas de reconhecimento de fala que permitem que as máquinas entendam e respondam à linguagem natural.
- Satélites e outros sensores espaciais que fornecem informações sobre o tempo e o clima.
- Sistemas de vigilância inteligente usando reconhecimento facial e detecção de objetos.
- Sensores de veículos autônomos, como câmeras, lidar, radar e sensores ultrassônicos.

A disponibilidade desses dados, aliada à presença de enormes recursos de armazenamento e de processamento computacional a um custo reduzido, impulsionou um maior apetite por modelagem avançada, manifestado no uso de uma ampla variedade de técnicas de machine learning e no desenvolvimento de inteligência artificial (AI) em praticamente todos os setores e âmbitos[17].

Embora haja consenso de que os modelos de AI geralmente oferecem maior poder de previsão do que os modelos tradicionais[18], teles também introduzem maior complexidade e podem ser difíceis de interpretar e explicar seus resultados.

Isso cria riscos associados ao uso desses modelos, como a falta de compreensão do modelo, a presença de vieses inadvertidos ou a dificuldade de determinar se o modelo está treinado em excesso (global ou localmente), o que pode levar a uma escassa capacidade de generalização e a possíveis erros nas decisões baseadas no modelo e, consequentemente, a uma falta de confiança no modelo.

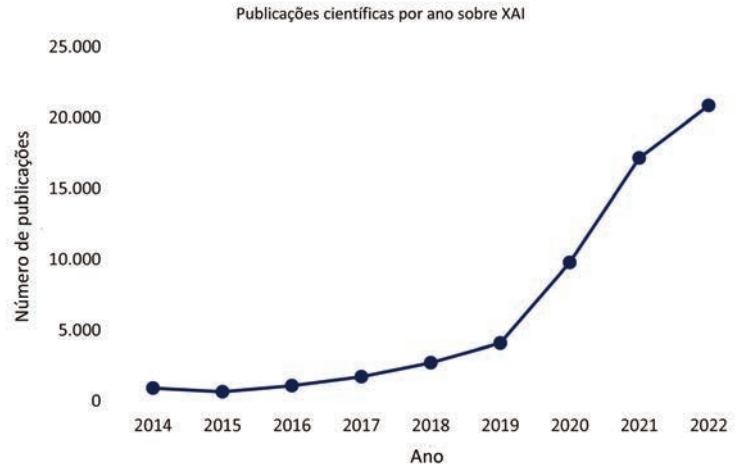
Isso levanta a questão de saber se é possível entender suficientemente bem os resultados dos algoritmos de AI, especialmente quando eles afetam decisões críticas, como diagnóstico médico, direção autônoma, detecção de fraudes e muitas outras.

[16] GPT-4, Generative Pre-Trained Transformer, uma rede neural profunda projetada pela OpenAI Foundation para executar tarefas de processamento de linguagem natural (NLP). Nesse caso, foi solicitado a ele que "apresentasse 10 citações inteligentes sobre inteligência artificial e quão difícil e necessário é ser capaz de interpretar e explicar os modelos de AI". A citação enviada foi a terceira.

[17] Embora existam diferenças, dada a falta de consenso sobre sua definição, os termos "aprendizado de máquina", "machine learning (ML)", "inteligência artificial (AI)" e "modelagem avançada" serão usados de forma intercambiável neste documento. Além disso, a abreviação "AI" será usada para "inteligência artificial", para fins de consistência com o acrônimo "XAI" (que geralmente não é traduzido), mesmo em citações de publicações em português.

[18] LeCun, Y. et al (2015). Pesquisador do Facebook AI Research e da Universidade de Nova York.

Figura 2. Número de publicações científicas por ano sobre Inteligência Artificial Explicável (XAI).



Definição

A disciplina de XAI é relativamente nova e, portanto, ainda não há uma doutrina estabelecida que padronize sua terminologia. Apesar de alguns esforços notáveis para definir os termos[19], a abordagem da XAI é heterogênea (dependendo da fonte acadêmica consultada) ou intuitiva (mais frequente na prática industrial).

De qualquer forma, para a maioria dos usos na prática, pode ser suficiente definir XAI da seguinte forma[20]:

> A inteligência artificial explicável (XAI) é o conjunto de processos e métodos que permite que os usuários humanos entendam e confiem nos resultados e produtos criados pelos algoritmos de machine learning. A XAI é usada para descrever um modelo de AI, seu impacto esperado e possíveis vieses. Ela ajuda a caracterizar a precisão, a imparcialidade, a transparência e os resultados do modelo na tomada de decisões baseada em AI. A XAI é fundamental para que uma organização crie confiança ao colocar modelos de AI em produção. A explicabilidade da AI também ajuda a organização a adotar uma abordagem responsável para o desenvolvimento da AI.

Relevância da XAI

Um aspecto sobre o qual há consenso entre acadêmicos e profissionais do setor é a crescente relevância da XAI como uma disciplina complementar à AI.

As ferramentas de análise de publicações científicas identificam mais de 77.000 artigos sobre XAI entre 2014 e 2022, e em uma tendência de aumento exponencial, com mais de 20.000 artigos somente em 2022 (Fig. 2)[21].

Além do interesse acadêmico, a atenção dada à XAI é explicada por sua capacidade de abordar uma série de preocupações do setor no uso da AI (Fig. 3), entre elas:

- **Requisitos regulatórios:** a obrigação de cumprir a regulamentação emergente sobre o uso de AI.
- **Falta de confiança:** a necessidade de criar confiança no modelo de AI e nos resultados que ele fornece entre os usuários, validadores e auditores e, em última análise, o público em geral.
- **Potencial uso indevido:** a conveniência de evitar o uso indevido dos modelos devido à falta de compreensão de como eles funcionam, o que pode resultar em custos e até mesmo em penalidades.
- **Impacto reputacional:** a prevenção de impactos reputacionais na empresa devido a preconceitos, decisões discriminatórias, uso inadequado ou simplesmente previsões errôneas do modelo.
- **Impactos sociais ou humanos:** a prevenção de impactos sociais ou humanos em usos críticos, como AI para diagnóstico de doenças médicas, decisões judiciais, identificação biométrica, polígrafos, etc.
- **Outros:** mitigação de outros riscos decorrentes da falta de compreensão do modelo, como segurança cibernética, proteção de dados, fraude, risco de modelo, etc.

Apesar de tudo isso, há casos em que os modelos de AI não precisam ser particularmente interpretáveis, porque os usos não estão regulamentados, porque não têm impactos potenciais relevantes ou simplesmente porque não precisam ser interpretados, como sistemas de recomendação automática de filmes e músicas ou algoritmos que jogam xadrez, por exemplo.

[19] Marcinkevics et al. (2020). Departamento de Ciência da Computação, ETH Zurich.
[20] IBM (2022).
[21] Dimensions (2022).

Figura 3. Preocupações do setor que a XAI está ajudando a resolver.

Regulamentação

Portanto, a XAI está se posicionando como uma disciplina de relevância crescente, o que está levando os órgãos reguladores e supervisores de diferentes jurisdições a estabelecer regulamentos e diretrizes para o uso adequado da AI, incluindo aspectos de interpretabilidade do modelo.

Nesse contexto, possivelmente as referências regulatórias mais relevantes no momento em que este artigo foi escrito são as seguintes:

1. GDPR (Parlamento Europeu)

Na Europa, o Regulamento Geral de Proteção de Dados, que entrou em vigor em 2018, estabelece o "direito a uma explicação" para os cidadãos, segundo o qual[22]:

> O titular dos dados deve ter o **direito de não estar sujeito a uma decisão,** que pode incluir uma medida, que avalie aspectos pessoais relacionados a ele, que se baseie exclusivamente no processamento automatizado e que produza efeitos legais sobre ele ou que o afete significativamente de forma semelhante, como a recusa automática de uma solicitação de crédito on-line ou de serviços de compras on-line em que não haja intervenção humana. [...]
>
> De qualquer forma, esse processamento deve estar sujeito a salvaguardas adequadas, que devem incluir informações específicas ao titular dos dados e o direito de obter intervenção humana, de expressar seu ponto de vista, de **receber uma explicação sobre a decisão tomada após essa avaliação** e de contestar a decisão.

Isso tem implicações críticas para o uso da AI e pode levar a questionamentos sobre sua viabilidade. Entretanto, nas palavras do Parlamento Europeu[23]:

> Certamente há uma tensão entre os princípios tradicionais de proteção de dados - limitação da finalidade, minimização de dados, tratamento especial de "dados sensíveis", limitação de decisões automatizadas - e a implantação total do poder da AI e do big data. Esses últimos envolvem a coleta de grandes quantidades de dados relacionados a indivíduos e suas relações sociais e seu processamento para fins que não foram totalmente determinados no momento da coleta. Entretanto, **há maneiras de interpretar, aplicar e desenvolver princípios de proteção de dados que sejam consistentes com os usos benéficos da AI e do big data.**

E isso está de acordo com o quarto princípio para o uso ético da AI estabelecido pelo Grupo de Alto Nível da Comissão Europeia sobre Inteligência Artificial[24]:

> **Explicabilidade:** os processos algorítmicos devem ser transparentes, os recursos e os objetivos dos sistemas de AI devem ser comunicados abertamente e as decisões devem ser explicadas às pessoas direta e indiretamente afetadas.

De qualquer forma, o GDPR tem impactos relevantes sobre o uso da AI, no sentido de que as empresas são legalmente obrigadas a explicar por que um modelo de AI produziu um determinado resultado, e isso tem implicações críticas para o desenho e a análise de interpretabilidade dos modelos de AI[25].

2. Artificial intelligence act (Parlamento Europeu)

O projeto de Regulamento de Inteligência Artificial ou Artificial Intelligence Act (AI Act), publicado em 2021, é uma proposta para o uso de inteligência artificial na União Europeia que visa garantir um alto nível de confiança na AI e em suas aplicações, ao mesmo tempo em que estabelece as bases para a inovação.

[22] GDPR (2018), Cons. 71.
[23] European Parliamentary Research Service (2020).
[24] Ibíd.
[25] Em alguns países europeus, o nível de conformidade desse tipo de AI (em especial os chamados Large Language Models) com a regulamentação de proteção de dados está sendo analisado e, em alguns casos, o uso de alguns desses modelos foi temporariamente proibido.

O regulamento estabelece um framework regulatório para sistemas de AI na UE e inclui requisitos de desenvolvimento ético, transparência, segurança e precisão. Ele também estabelece um sistema de governança e supervisão para sistemas de AI, bem como regras de proteção e governança de dados.

Sendo um regulamento, quando for adotado, será de aplicação direta nos 27 países da União[26], sem a necessidade de ser transposto para a legislação de cada país.

Uma de suas principais características é que ele classifica os aplicativos de AI em níveis de risco[27]:

- **Práticas proibidas,** que denotam a categoria de maior risco; esses sistemas são totalmente proibidos. Eles incluem:
  - Sistemas biométricos em tempo real que podem ser usados para qualquer tipo de vigilância, embora haja exceções para a prevenção de crimes e investigações criminais em contextos de aplicação da lei e segurança nacional.
  - Algoritmos de pontuação social que podem ser usados para avaliar indivíduos com base em características pessoais ou comportamento de uma forma que possa causar danos ou levar a um tratamento desfavorável de um indivíduo.
  - Sistemas manipuladores que exploram as vulnerabilidades de indivíduos específicos para distorcer seu comportamento de forma que possa causar danos físicos ou psicológicos.

- **Sistemas de AI de alto risco,** que estão listados no Anexo III e provavelmente constituirão a maioria dos sistemas de AI. Esses sistemas incluem:
  - Identificação biométrica e categorização de pessoas físicas [...].
  - Gestão e operação de infraestrutura crítica [...] [por exemplo, tráfego].
  - Educação e formação profissional [...].
  - Emprego e gestão de trabalhadores [...].
  - Acesso a serviços essenciais [...], incluindo a avaliação da capacidade de crédito, classificação de crédito ou priorização do acesso a esses serviços (Observação: isso se aplica especialmente aos sistemas de AI usados no setor de serviços financeiros).
  - Forças de segurança [...].
  - Gerenciamento de controles de fronteira [...].
  - Administração da justiça e processos democráticos [...].

- **Sistemas de AI de baixo risco [ou risco limitado],** que incluem sistemas que não usam dados pessoais ou fazem previsões que possam afetar direta ou indiretamente qualquer indivíduo, como aplicativos de manutenção preditiva industrial.

Com relação à interpretabilidade dos modelos de AI classificados como de alto risco, a Lei de AI estabelece[28] em seus artigos 13 e 14:

Art. 13. Transparência e comunicação de informações aos usuários

1. Os sistemas de AI de alto risco devem ser projetados e desenvolvidos de forma a garantir que operem com um nível suficiente de transparência para que suas informações de saída sejam corretamente interpretadas e utilizadas pelos usuários. [...]

2. Os sistemas de AI de alto risco devem ser acompanhados de instruções apropriadas para uso em formato digital ou outro formato apropriado, que devem incluir informações concisas, completas, corretas e claras que sejam relevantes, acessíveis e compreensíveis para os usuários. [...]

Art. 14. Vigilância humana

1. Os sistemas de AI de alto risco devem ser projetados e desenvolvidos de forma que possam ser efetivamente monitorados por pessoas físicas durante o período em que estiverem em uso, inclusive fornecendo-lhes uma ferramenta de interface homem-máquina adequada, entre outras coisas.

[...]

4. As medidas acima [...] devem permitir que as pessoas encarregadas da supervisão humana sejam capazes, dependendo das circunstâncias:
   a. **Compreender totalmente os recursos e as limitações do sistema de AI de alto risco** e controlar adequadamente seu funcionamento, de modo que possam detectar sinais de anomalias, mau funcionamento e comportamento inesperado e resolvê-los o mais rápido possível;
   b. estar ciente da possível tendência de confiar automaticamente ou em excesso nas informações de saída geradas por um sistema de AI de alto risco ("viés de automação") [...];
   c. interpretar corretamente as informações de saída do sistema de AI de alto risco [...];
   d. decidir, em uma determinada situação, não usar o sistema de AI de alto risco ou desconsiderar, invalidar ou reverter as informações de saída geradas por ele;
   e. intervir na operação do sistema de AI de alto risco ou interromper o sistema [...].

Como pode ser visto, o AI Act impõe condições restritivas sobre a interpretabilidade dos modelos de AI de alto risco (Fig. 4), que

[26]Espera-se que ela entre em vigor 20 dias após sua publicação no Diário Oficial da União Europeia e que seja de plena aplicação 24 meses após sua entrada em vigor.
[27]Floridi et al. (2022).
[28]Comissão Europeia (2021).

logo se tornarão obrigatórios em toda a União. Espera-se que isso desencadeie um número significativo de iniciativas de adaptação ao Regulamento, incluindo uma documentação mais abrangente dos modelos e seus usos, a aplicação de técnicas de interpretabilidade, o desenvolvimento de dashboards de monitoramento e de alerta de modelos ou a revisão do procedimento integrado para desenvolvimento, validação, implementação e uso de modelos, entre outros.

3. Diretrizes éticas para uma Inteligência Artificial confiável (Comissão Europeia)

Em abril de 2019, o Grupo de Especialistas de Alto Nível sobre AI da Comissão Europeia apresentou as Diretrizes Éticas para uma AI Confiável[29], após um processo de consulta com mais de 500 respostas do setor.

As Diretrizes propõem sete requisitos principais que os sistemas de AI devem atender para serem considerados confiáveis, que em resumo são: (i) ação humana e supervisão, (ii) solidez técnica e segurança, (iii) privacidade e gestão de dados, (iv) transparência, (v) diversidade, não discriminação e equidade, (vi) bem-estar ambiental e social e (vii) responsabilização.

Em particular, no que diz respeito à interpretabilidade dos modelos de AI, as Diretrizes declaram o seguinte como parte de seu requisito de transparência:

> 53. A explicabilidade é fundamental para conquistar e manter a confiança dos usuários nos sistemas de AI. Isso significa que os processos precisam ser transparentes, que os recursos e a finalidade dos sistemas de AI precisam ser comunicados abertamente e que **as decisões devem, na medida do possível, poder ser explicadas às partes que são** direta ou indiretamente **afetadas por elas.** Sem essas informações, não é possível contestar adequadamente uma decisão.
>
> **Nem sempre é possível explicar por que um modelo gerou um determinado resultado ou decisão** (ou qual combinação de fatores contribuiu para isso). Esses casos, que são chamados de algoritmos de "black box", exigem atenção especial.
>
> Em tais circunstâncias, **outras medidas relacionadas à explicabilidade** (por exemplo, rastreabilidade, auditabilidade e comunicação transparente sobre o desempenho do sistema) **podem ser necessárias, desde** que o sistema como um todo respeite os direitos fundamentais.
>
> **O grau de necessidade de explicabilidade depende, em grande parte, do contexto e da gravidade das consequências** de um resultado errôneo ou inadequado.

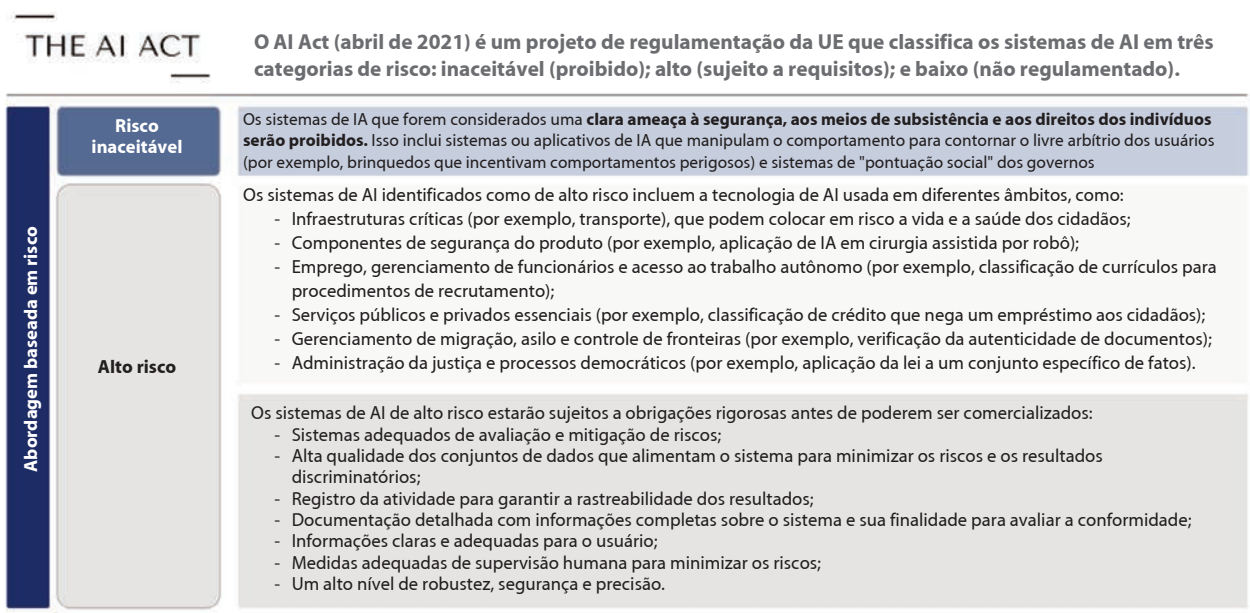
Como se pode ver, as Diretrizes apontam para a mesma direção: a exigência (que se eleva ao nível de necessidade ética) de que os modelos de AI sejam explicáveis.

Além disso, o que à primeira vista pode parecer um requisito mais relaxado para a interpretabilidade dos modelos de AI, uma vez que as Diretrizes reconhecem que alguns modelos de AI são mais difíceis de explicar, na verdade introduz uma complexidade adicional: a necessidade de classificar os modelos de AI de acordo com seu risco e seu potencial de interpretação, a fim de aplicar um grau maior ou menor de esforço em sua explicação.

Por fim, as Diretrizes têm como objetivo avaliar até que ponto um modelo de AI atende a esses sete requisitos, propondo uma lista de critérios de avaliação, que deve ser adaptada a cada caso específico. Com relação à explicabilidade, as Diretrizes formulam os seguintes critérios de avaliação[30], que devem ser integrados a outras ferramentas de avaliação já disponíveis para as organizações:

[29] Comissão Europeia (2019).
[30] Ibíd

Figura 4. Áreas de aplicação e requisitos do Artificial Intelligence Act.



- Você avaliou até que ponto as decisões e, portanto, o resultado produzido pelo sistema de AI são compreensíveis?
- Foi assegurado que é possível desenvolver uma explicação que seja compreensível para todos os usuários que desejam saber por que um sistema tomou uma decisão específica que levou a um resultado específico?
- Você avaliou até que ponto a decisão do sistema influencia os processos de tomada de decisão da organização?
- Você avaliou por que esse sistema específico foi implantado nessa área específica?
- Você avaliou o modelo de negócios do sistema (por exemplo, como ele cria valor para a organização)?
- Você desenhou o sistema de AI tendo em mente a interpretabilidade desde o início?
- Você pesquisou e tentou usar o modelo mais simples e mais interpretável possível para a aplicação em questão?
- Você já avaliou se pode analisar seus dados de treinamento e teste e se pode modificar e atualizar esses dados ao longo do tempo?
- Você avaliou se, após o treinamento e o desenvolvimento do modelo, tem alguma possibilidade de revisar sua interpretabilidade ou se tem acesso ao fluxo de trabalho interno do modelo.

4. Blueprint for an AI Bill of Rights (Casa Branca)

Em outubro de 2022, a Casa Branca propôs uma minuta da Declaração de Direitos sobre Inteligência Artificial[31], promovida pelo presidente Joe Biden e desenvolvida pelo Escritório de Política de Ciência e Tecnologia da Casa Branca (OSTP), e acompanhada de um manual (From Principles to Practice) sobre como implementá-la na prática.

O AI Bill of Rights estabelece cinco princípios ou direitos dos cidadãos em relação ao AI, que estão resumidos em[32]:

- Sistemas seguros e eficazes.
- Proteção contra discriminação de algoritmos.
- Privacidade de dados.
- Notificação e explicação.
- Processo alternativo de avaliação e correção humana em caso de falha de AI (fallback).

Em seu quarto princípio, referente à explicabilidade dos modelos de AI, ele afirma, entre outros, que[33]:

> Os projetistas, desenvolvedores e implementadores de sistemas automatizados devem fornecer documentação em linguagem simples e geralmente acessível, que inclua descrições claras da operação geral do sistema. [...]
>
> Os sistemas automatizados devem ser acompanhados de explicações que sejam tecnicamente válidas, significativas e úteis para você e para qualquer operador ou outras pessoas que precisem entender o sistema. [...]
>
> Os sistemas automatizados devem fornecer notificações de uso comprovadamente claras, oportunas, compreensíveis e acessíveis, além de **explicações sobre como e por que o sistema tomou uma decisão** ou executou uma ação.

5. Princípios sobre Inteligência Artificial (OECD)

Os Princípios da OCDE sobre Inteligência Artificial promovem o uso de AI que seja confiável e respeite os direitos humanos e os valores democráticos. Eles foram adotados em maio de 2019 pelos 38 países membros da OCDE. Foram os primeiros princípios desse tipo a serem endossados pelos governos e incluem recomendações concretas para políticas públicas e estratégias sobre AI.

Entre outros, eles afirmam que "os responsáveis da AI devem se comprometer com a transparência e a divulgação responsável dos sistemas de AI. Para esse fim, eles devem fornecer informações significativas, adequadas ao contexto e consistentes com o estado da técnica [...] para que aqueles afetados por um sistema de AI possam entender o resultado"[34]. O Observatório de Políticas de AI da OCDE, lançado em fevereiro de 2020, tem como objetivo ajudar os tomadores de decisão a implementar esses Princípios.

6. Discussion Paper on Machine Learning for IRB Models (EBA)

O Discussion Paper on Machine Learning for IRB Models, da Autoridade Bancária Europeia (EBA), publicado em novembro de 2021, é particularmente relevante para o setor bancário (Fig. 5).

O documento tem como objetivo analisar a relevância dos possíveis obstáculos à implementação de técnicas de machine learning no contexto da abordagem IRB para o cálculo de capital em instituições financeiras, inclui os desafios e os possíveis benefícios do uso dessas técnicas e estabelece determinados princípios e recomendações[35]. Um foco central do documento é, obviamente, como tornar o uso dessas técnicas compatível com a conformidade com o Regulamento de Capital Europeu (CRR[36]).

[31] OSTP da Casa Branca (2022).
[32] Ibíd.
[33] Ibíd.
[34] OECD (2019).
[35] Ver análise detalhada na Management Solutions (2021).
[36] CRR: Capital Requirements Rule (Regra de Requisitos de Capital), regulamentação central sobre capital em instituições financeiras na Europa.

Com relação à interpretabilidade dos modelos, o documento aborda essa questão sob o título "Concerns about the use of machine learning techniques" (Preocupações sobre o uso de técnicas de machine learning) e afirma[37]:

> As principais preocupações decorrentes da análise dos requisitos da CRR estão relacionadas à complexidade e à confiabilidade dos modelos de ML, em que os principais desafios parecem ser a interpretabilidade dos resultados, a governança, com referência específica ao aumento das necessidades de formação do pessoal, e a dificuldade de avaliar a capacidade de generalização de um modelo (ou seja, evitar o overfitting).
>
> Para entender as relações subjacentes entre as variáveis exploradas pelo modelo, os profissionais desenvolveram várias técnicas de interpretabilidade [...] [e] a escolha de qual dessas técnicas usar pode representar um desafio em si, pois essas técnicas geralmente permitem apenas uma compreensão limitada da lógica do modelo.

Além disso, o documento introduz a necessidade de se encontrar um equilíbrio entre a complexidade e a interpretabilidade do modelo e, diferentemente de outras regulamentações, desce a um nível mais técnico ao recomendar às instituições financeiras:

a. Analisar de forma estatística: i) a relação de cada variável de entrada com a variável de saída, ceteris paribus; ii) o peso global de cada variável de entrada na determinação da variável de saída, para detectar quais variáveis têm maior influência na predição do modelo. Estas análises são particularmente relevantes quando não é possível determinar uma representação próxima e pontual da relação entre a variável de saída do modelo e as variáveis de entrada devido à complexidade do modelo.

b. Avaliar a relação econômica de cada variável de entrada com a variável de saída para garantir que as estimativas do modelo sejam plausíveis e intuitivas.

c. Apresentar um documento de síntese que explique de forma simples o modelo baseado nos resultados das análises descritas na alínea a. O documento deve, no mínimo, descrever:

   i. Os principais fatores do modelo.
   ii. As principais relações entre as variáveis de entrada e as previsões do modelo.

   O documento é dirigido a todas as partes interessadas, incluindo a equipe que usa o modelo para fins internos.

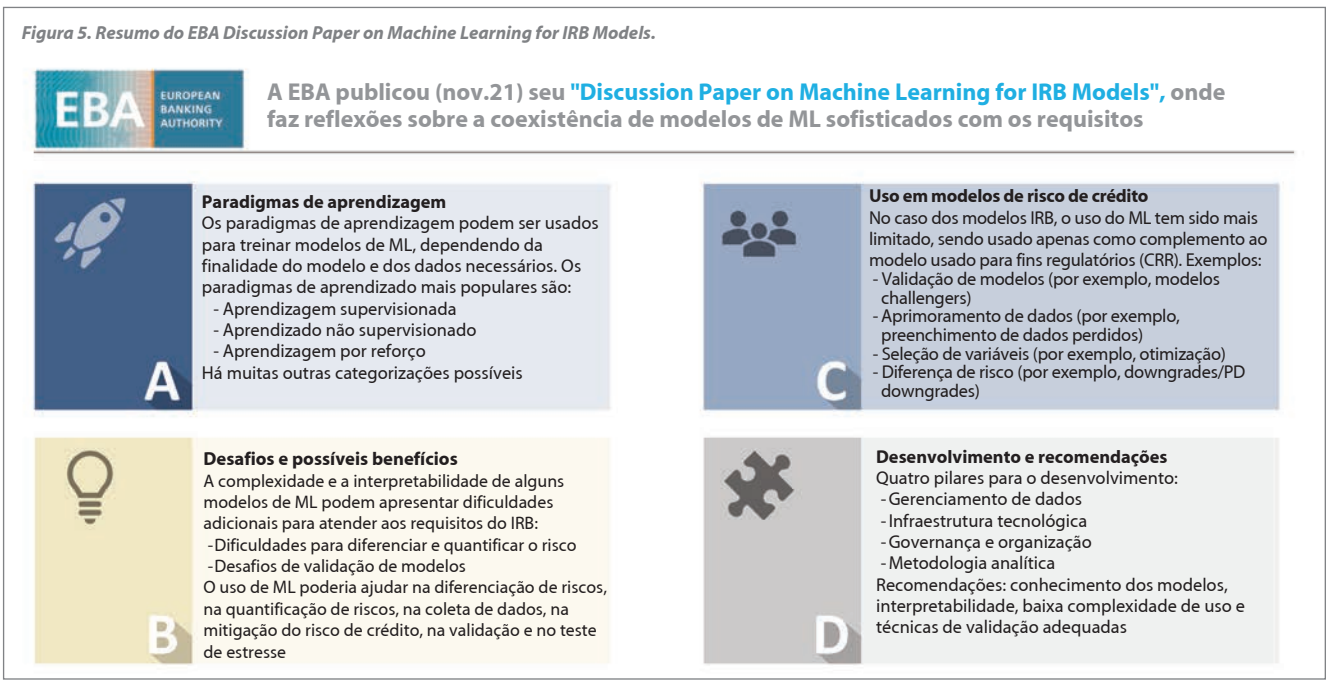
d. Garantir a detecção de possíveis vieses no modelo (por exemplo, overfitting à amostra de treinamento).

Na prática, enquanto o setor bancário aguarda a versão final do documento consultivo da EBA, a maioria das instituições que usam técnicas de machine learning em seus modelos IRB já está adaptando suas estruturas de desenvolvimento, monitoramento e validação de modelos para garantir a conformidade futura.

Um elemento comum em todas as referências regulatórias mencionadas acima, como pode ser visto, é a necessidade de fornecer uma explicação aos cidadãos sobre o uso da AI, e fazê-lo em dois níveis: a interpretabilidade e a transparência do modelo de AI como um todo, e a capacidade de explicar uma decisão específica do modelo, se necessário.

[37] EBA (2021).

Figura 5. Resumo do EBA Discussion Paper on Machine Learning for IRB Models.



Além das referências regulatórias descritas acima, há um grande número de publicações, princípios, diretrizes e projetos de regulamentação em várias jurisdições que abordam a interpretabilidade dos modelos de AI, tanto gerais quanto setoriais, e tanto regionais quanto locais em cada país; a seleção apresentada nesta seção inclui aqueles considerados como tendo o maior escopo e influência potencial.

Impactos na organização e nos processos

Um princípio essencial da XAI como disciplina é que, além do desenvolvimento de técnicas específicas de explicabilidade ou da construção de modelos inerentemente interpretáveis, essa explicabilidade e interpretabilidade devem ser integradas à organização e aos processos da empresa.

Colocado em prática, esse princípio implica o desenvolvimento e a implementação de um framework de XAI, que pode ser estruturado em quatro elementos:

1. Técnicas de interpretabilidade de modelos de AI
2. Integração aos processos de gestão de risco de modelo (MRM)
3. Suporte tecnológico
4. Fator humano

1. Técnicas de interpretabilidade dos modelos de AI

No centro de um framework de XAI estão as técnicas de interpretabilidade e explicabilidade, que podem ser resumidas em três aspectos:

- **Interpretabilidade do desenho do modelo:** isso inclui analisar como o modelo se comportaria em diferentes cenários (por exemplo, ataques adversários, cenários extremos...), entender como os submodelos e os conjuntos de modelos funcionam e integrar a interpretabilidade ao desenho do modelo aplicando restrições durante o desenvolvimento do modelo.

- **Interpretabilidade dos resultados do modelo:** refere-se à detecção de quais variáveis influenciam a previsão do modelo e como, por meio da interpretabilidade local (LIME, SHAP, etc.) e global (PDP, significância da variável, modelos substitutos, análise de sensibilidade); à avaliação do sentido econômico de cada variável (por exemplo, análise de caso de uso de uma amostra representativa de dados); e à garantia de que a documentação do modelo descreva corretamente o modelo, incluindo as variáveis de entrada e sua relação com os resultados.

- **Outros aspectos:** garantir a detecção de possíveis vieses no modelo (por exemplo, overfitting, dados de entrada tendenciosos, erros de dados) e monitorar regularmente o modelo, especialmente quando seu escopo mudar ou quando for aplicado a dados diferentes dos dados de desenvolvimento.

Devido à sua importância, as principais técnicas de interpretabilidade e explicabilidade serão desenvolvidas na seção a seguir.

2. Integração aos processos de gestão do risco de modelo (MRM)

A interpretabilidade dos modelos de AI é uma característica que transcende o desenvolvimento e afeta toda a cadeia do ciclo de vida do modelo e, portanto, todo o framework de gestão de risco do modelo. Um resumo não exaustivo da incorporação da XAI ao framework de MRM de uma empresa inclui a análise dos seguintes elementos:

- **Governança:** atualizar o framework de organização e de governança para incorporar a XAI; avaliar o impacto da regulamentação aplicável aos modelos de AI; atualizar o sistema de classificação de modelos para abordar a falta de interpretabilidade como um risco importante; atualizar o inventário de modelos e os procedimentos de inventário para incorporar elementos da XAI (por exemplo, atributos específicos para modelos de AI).

- **Desenvolvimento:** atualizar as políticas e os procedimentos de desenvolvimento de modelos, bem como os requisitos de documentação; avalie a imparcialidade e a parcialidade, a interpretabilidade das entradas, o design e as saídas, os dados, o risco de fornecedores, as métricas de capacidade preditiva, os limites para o uso de modelos de AI etc.; realizar a análise de sensibilidade dos modelos de AI para identificar vulnerabilidades; incluir no framework de desenvolvimento testes específicos para XAI.

- **Monitoramento:** atualizar o framework de monitoramento de modelos e completá-la com testes específicos de XAI; revisar limites e ações para não conformidade; desenvolver sistemas de alerta antecipado para detectar mudanças nos modelos de AI; revisar a conformidade com o apetite de risco de modelo; avaliar a necessidade de desenvolver um

módulo de monitoramento ad hoc para modelos de aprendizado dinâmico (ou seja, modelos que se recalibram automaticamente sem intervenção humana).

- **Validação:** atualizar o framework de validação interna para detectar possíveis riscos associados aos modelos de AI e incorporar testes de XAI; estabelecer um framework de validação cruzada para garantir a qualidade dos modelos de AI; avaliar o impacto das mudanças no ambiente de produção nos modelos de AI.

- **Implementação:** atualizar o processo de implementação do modelo para incorporar testes específicos às características da XAI; atualizar, quando apropriado, a plataforma tecnológica para permitir a produção de modelos de AI.

- **Uso:** atualizar procedimentos para o uso de modelos de AI para determinar sua adequação ao contexto em que serão usados; revisar e concluir o treinamento de usuários em modelos de AI; atualizar protocolos para detectar possíveis situações de uso indevido ou exploração de modelos.

- **Auditoria:** implementar um framework de auditoria para modelos de AI para garantir sua implementação e uso adequados; estabelecer testes de XAI para a auditoria de modelos de AI; avaliar a adequação dos sistemas de controle interno para garantir a qualidade dos modelos de AI; analisar trilhas de auditoria para detectar possíveis riscos associados aos modelos de AI.

Portanto, o uso de modelos de AI implica uma revisão completa das políticas e dos procedimentos durante todo o ciclo de vida do modelo para incorporar, no mínimo, os elementos da XAI.

3. Suporte tecnológico

A implementação de um framework de XAI tende a começar com ferramentas departamentais e, assim que atinge um nível mínimo de maturidade, requer soluções tecnológicas profissionais para dar suporte aos aspectos de interpretabilidade dos modelos de AI.

Essas soluções podem ser classificadas em dois grupos:

- Interpretabilidade: desenvolvimento de sistemas que implementem técnicas de interpretabilidade de forma padronizada e homogênea. Eles devem permitir que a interpretação dos modelos seja realizada automaticamente, facilmente configurável e com alta qualidade, incorporando as técnicas mais comuns e oferecendo flexibilidade para adicionar novas técnicas à medida que forem desenvolvidas[38].

- Governança de modelos: desenvolvimento ou atualização de sistemas de governança de modelos para dar suporte aos aspectos de XAI em MRM (inventário, classificação, documentação etc.), garantindo assim que os modelos disponíveis atendam aos requisitos de qualidade, segurança e explicabilidade exigidos[39].

Além disso, recomenda-se uma abordagem holística que englobe todos os aspectos do framework de XAI. Isso inclui o uso de ferramentas de análise de dados, o desenvolvimento de APIs para a integração dos sistemas de interpretabilidade e governança de modelos descritos acima, a criação de mecanismos de segurança e auditoria e a definição de protocolos para garantir a conformidade com os padrões de qualidade e explicabilidade.

4. Fator humano

Um quarto elemento na integração da XAI à organização e aos processos é a consideração do fator humano. Isso inclui, entre outros:

- **Recrutamento e retenção de talentos:** desenvolvimento de programas de recrutamento e retenção de talentos especializados em XAI, para garantir a presença de profissionais com o conhecimento técnico e a experiência necessários para aplicar XAI na empresa, o que é especialmente relevante em um mercado de trabalho com escassez desse perfil profissional.

- **Treinamento:** desenvolvimento de programas de treinamento para equipes de desenvolvimento de modelos de AI, equipes de governança de modelos e usuários de modelos de AI para garantir que todos os envolvidos entendam os princípios básicos da XAI e como aplicá-los no contexto específico da empresa.

- **Cultura:** desenvolver uma cultura empresarial que incentive o uso e a exploração da explicabilidade e interpretabilidade dos modelos de AI. Isso pode incluir a adoção de metodologias ágeis para o desenvolvimento de modelos de AI, a criação de uma cultura de colaboração entre as equipes de desenvolvimento e governança de modelos e a consideração da explicabilidade como um fator crítico na aprovação de modelos de AI.

- **Gestão de mudanças:** desenvolvimento de programas de gestão de mudanças para garantir a adoção adequada da XAI pelas equipes da empresa que trabalham com modelos de AI. Isso inclui a motivação das equipes de desenvolvimento, a análise dos custos e benefícios da explicabilidade, a definição de protocolos de comunicação com terceiros, etc.

Em conclusão, a explicabilidade e a interpretabilidade dos modelos de AI são aspectos fundamentais que precisam ser integrados à organização e aos processos da empresa por meio de um framework de XAI adequado e abrangente, o que é essencial para garantir o uso desses modelos de acordo com a regulamentação e as boas práticas.

[38] Nesse sentido, a Management Solutions tem o ModelCraft™, um sistema proprietário de modelagem de componentes e AutoML, que incorpora um módulo completo de interpretabilidade. Ver Management Solutions (2023).

[39] A Management Solutions também possui o Gamma™, um sistema proprietário de governança de modelos que abrange todos os aspectos acima. Ver Management Solutions (2022).

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:

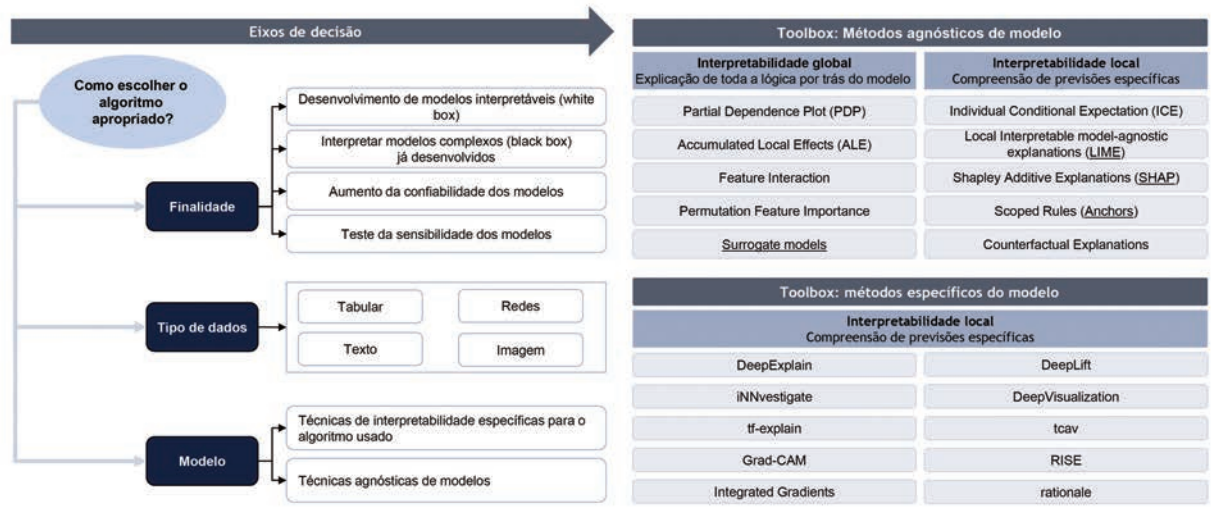
- A reprodutibilidade dos resultados, o processo de treinamento e implementação do modelo[47],a consistência de suas previsões e a explicação da sequência de previsões mais prováveis.
- Potencial de viés[48] nos dados de entrada.
- Imparcialidade (*fairness*)[49].
- Precisão da explicação[50].
- Solidez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

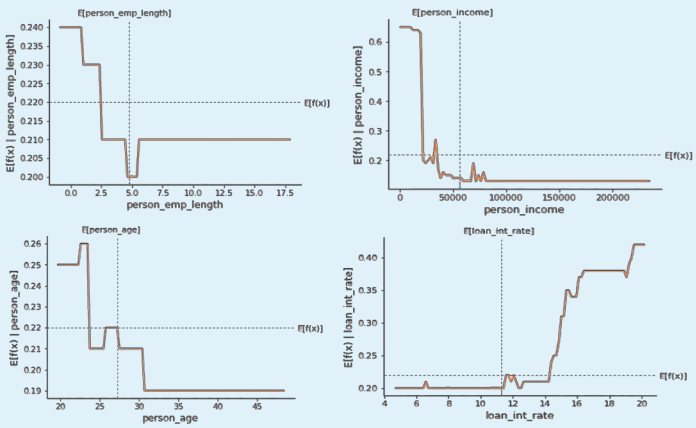
Os PDPs podem ser aplicados a um caso de uso muito comum no setor bancário: a pontuação dos clientes durante o processo de concessão de empréstimo para determinar a probabilidade de inadimplência. Neste exemplo, foi usada uma carteira anônima de empréstimos imobiliários com informações sobre sua atividade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

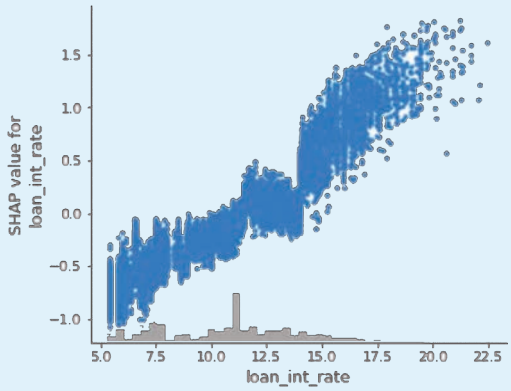
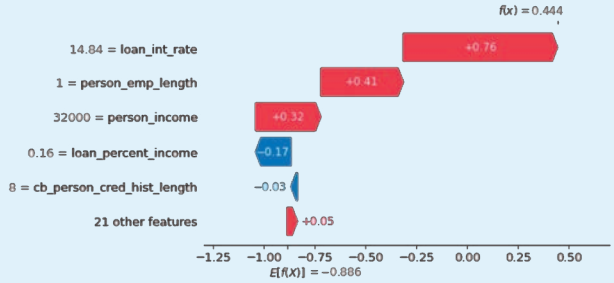
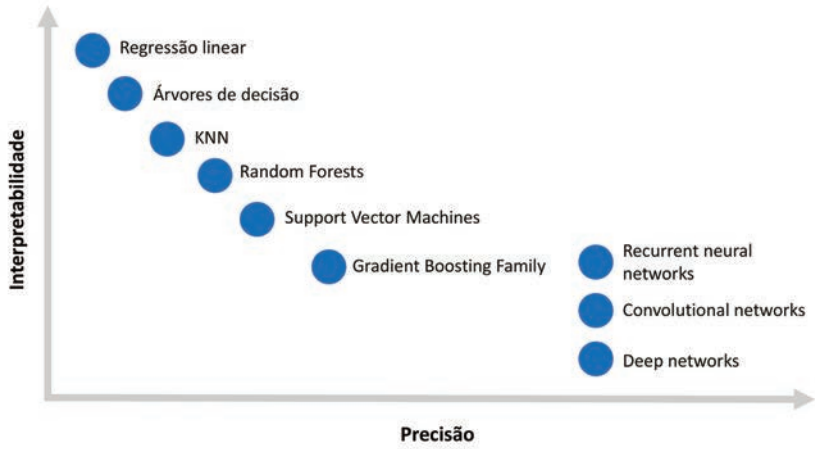


Figura 9. Valores de Shapley que influenciam a previsão de um cliente com um empréstimo negado[2].



[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

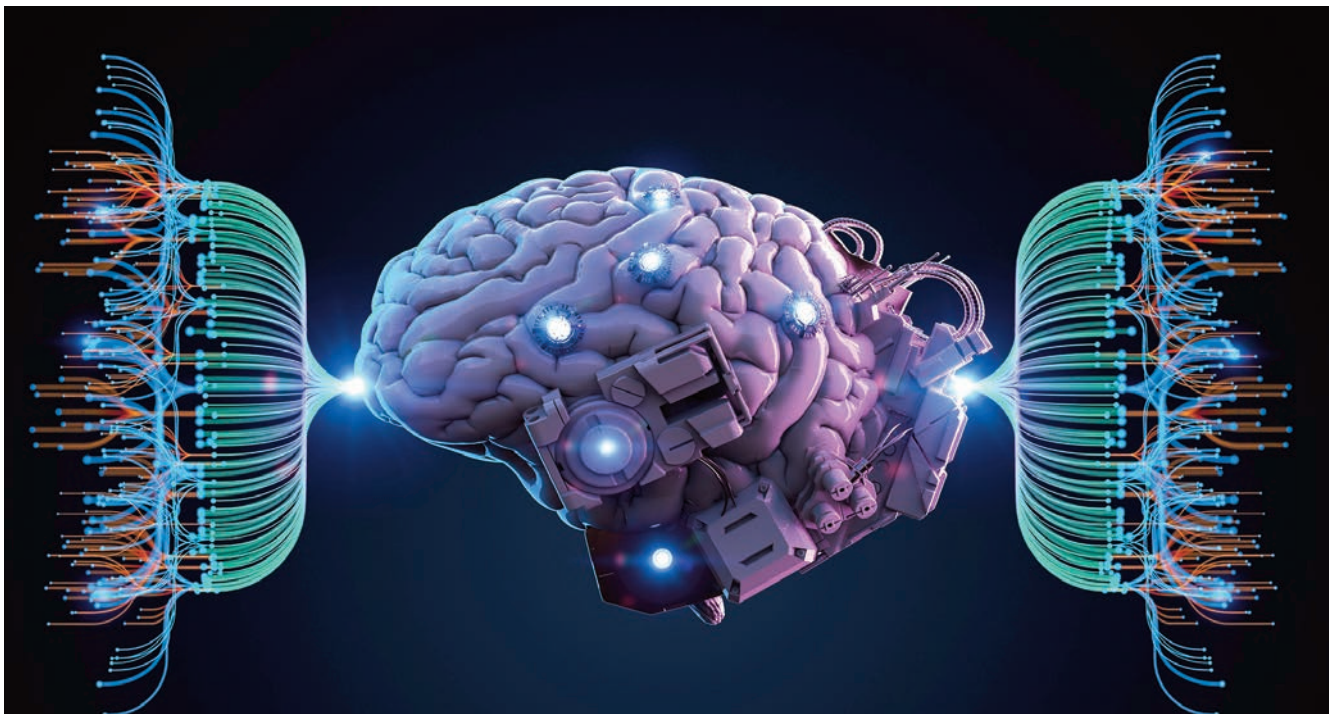
Em combinação com os desafios mostrados nesta seção, há outros elementos-chave que podem ser considerados para melhorar a interpretabilidade do modelo, como a imparcialidade do modelo, a ausência de viés nos dados de entrada, componentes especializados em potencial ou desempenho adequado e estrutura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

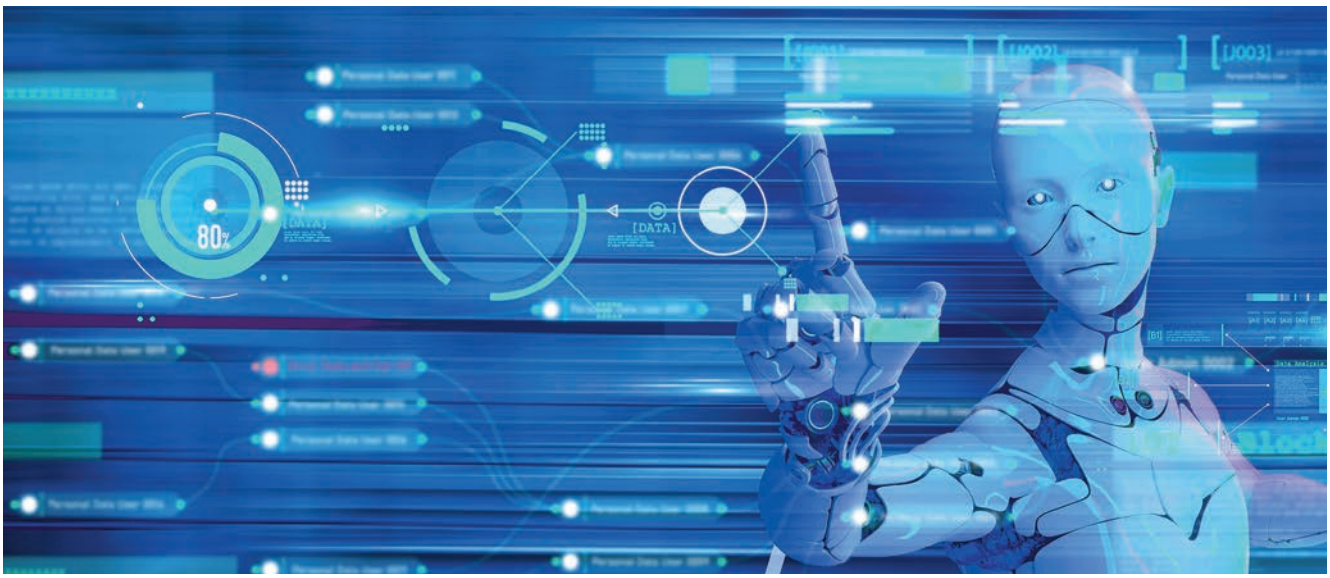
Em primeiro lugar, a interpretabilidade dos modelos ainda é limitada por vários fatores, como a reprodutibilidade dos resultados[70], o processo de treinamento e implementação do modelo, a consistência de suas previsões, a explicação da sequência de previsões mais prováveis, os vieses nos dados de entrada,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

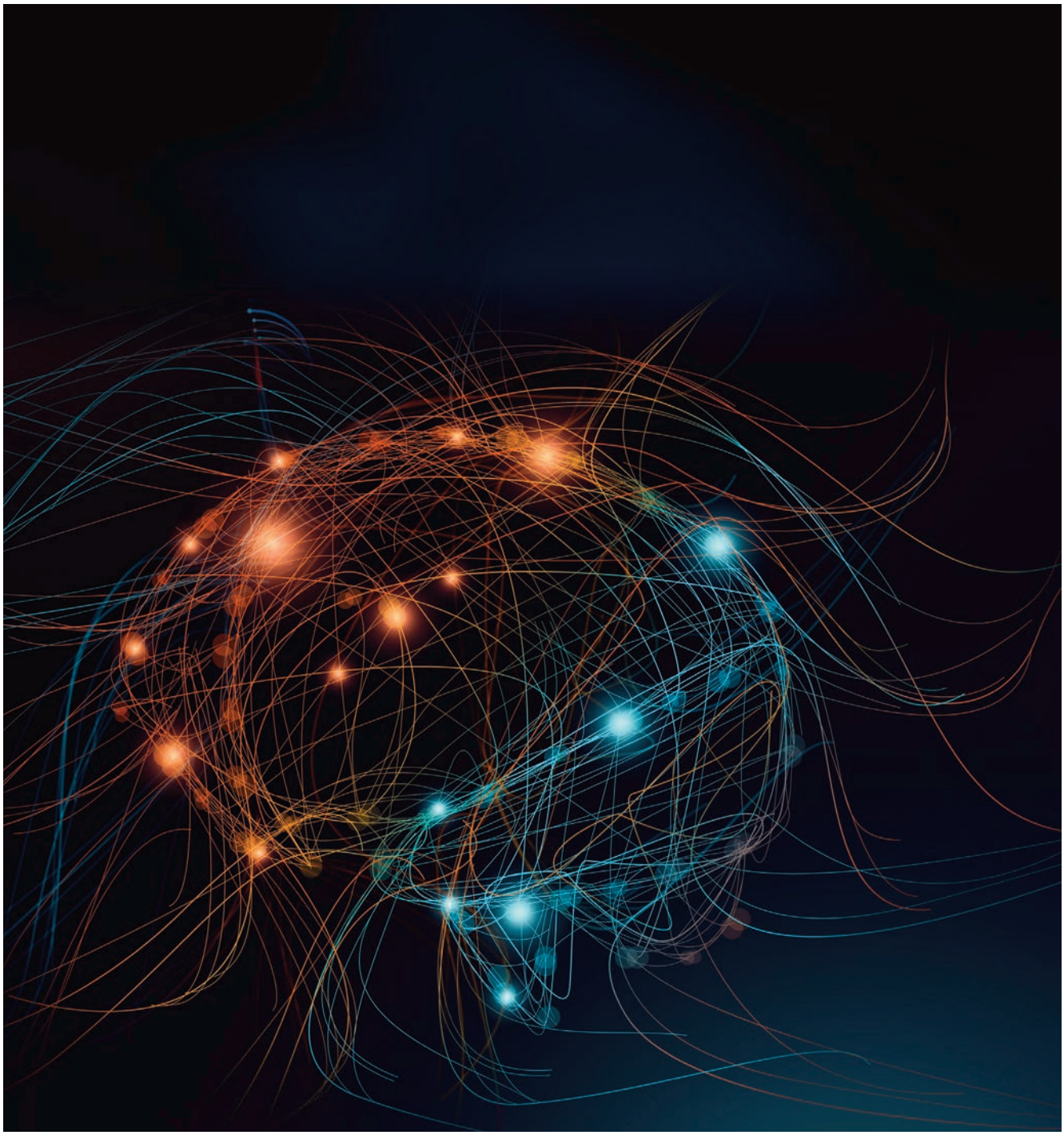
Nesse sentido, novas técnicas estão sendo desenvolvidas para melhorar a interpretabilidade dos modelos, como o uso de informações das camadas intermediárias de redes neurais profundas, a agregação de métricas de interpretabilidade para medir a explicabilidade dos modelos, o desenvolvimento de modelos 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.

Estudo de caso de interpretabilidade

«Os tolos ignoram a complexidade. Os pragmáticos sofrem com ela. Alguns conseguem evitá-la. Os gênios a eliminam»
Alan Perlis[72]



Abordagem

Esta seção apresenta um estudo de caso de interpretabilidade em inteligência artificial para ilustrar como as técnicas de XAI descritas na seção anterior são aplicadas.

O estudo de caso selecionado aborda o problema da retenção de funcionários em uma organização, concentrando-se em entender e explicar as causas que levam os funcionários a deixar seus empregos. A identificação desses fatores pode permitir que as organizações tomem medidas preventivas e desenvolvam estratégias para melhorar a satisfação no trabalho e a retenção de talentos.

Neste estudo de caso, será usado um conjunto de dados fictício gerado pela IBM e publicado no Kaggle[73]. Esse conjunto de dados contém informações sobre os funcionários de uma organização, incluindo características demográficas, detalhes do cargo e se eles deixaram ou não a empresa.

No exercício atual, a empresa tem uma taxa de fuga de funcionários de 16%, 6% acima da média histórica, e está preocupada em entender as causas para desenvolver um plano de remediação.

As principais variáveis presentes no conjunto de dados incluem:

- Nível de educação (de "secundário" a "doutorado")
- Satisfação com o ambiente de trabalho (de "baixa" a "muito alta")
- Envolvimento no trabalho (de "baixo" a "muito alto")
- Satisfação no trabalho (de "baixa" a "muito alta")
- Classificação de desempenho (de "baixo" a "excelente")
- Satisfação com as relações de trabalho (de "baixa" a "muito alta")
- Equilíbrio entre vida pessoal e profissional (de "ruim" a "ótimo")
- Anos desde a última promoção no emprego (variável numérica)
- Salário mensal (variável numérica)
- Anos no emprego atual (variável numérica)
- Distância até a estação de trabalho (variável numérica)
- Número de empresas nas quais o trabalho foi realizado (variável numérica)
- Cargo atual (variável categórica, inclui "Gerente", "Diretor", "Research Scientist", etc.)

O foco do estudo de caso será treinar e validar diferentes modelos de inteligência artificial para prever o desgaste dos funcionários, usando técnicas de XAI para analisar e entender o comportamento e as decisões dos modelos selecionados.

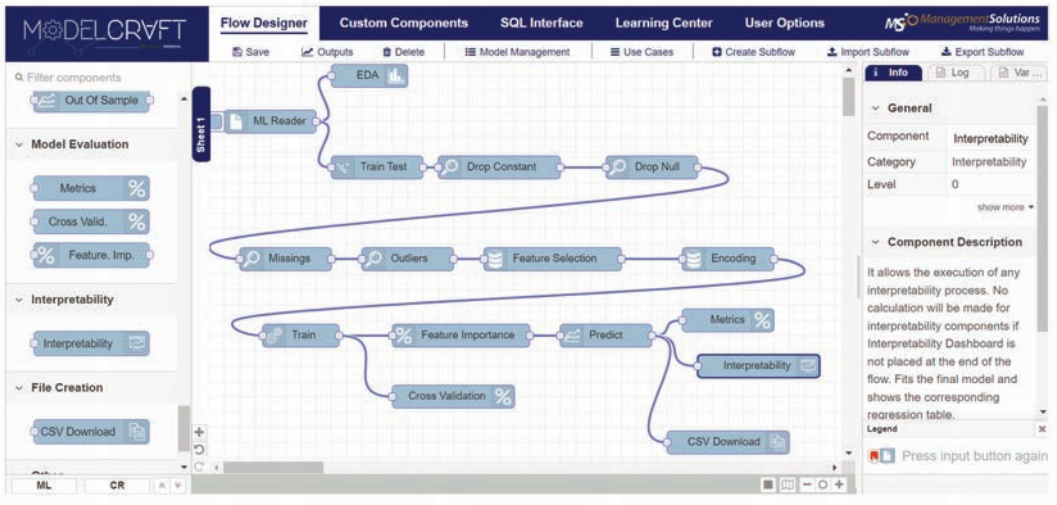
Para simplificar e acelerar o processo, foi usado o sistema de modelagem de componentes ModelCraft™, que contém várias técnicas relevantes de AI e XAI. Esse sistema permitirá que o estudo seja realizado de forma eficiente e sem a necessidade de escrever código.

Ao longo do estudo de caso, as técnicas de interpretabilidade SHAP, LIME e PDP serão aplicadas para analisar os modelos selecionados e entender quais variáveis influenciam as decisões dos funcionários de deixar seus empregos. Além disso, exploraremos como essas variáveis interagem umas com as outras e como elas afetam diferentes segmentos da população de funcionários.

[72] Alan Jay Perlis (1922-1990), cientista da computação americano, PhD em Ciência da Computação pelo MIT e professor da Universidade Purdue, da Universidade Carnegie Mellon e da Universidade da Califórnia em Berkeley, conhecido por seu trabalho pioneiro em linguagens de programação e por ser o primeiro ganhador do Prêmio Turing.

[73] Kaggle (2017). IBM HR Analytics Employee Attrition & Performance.

Figura 13. Fluxo de modelagem no ModelCraft™.



Ao final do estudo de caso, serão avaliadas a eficácia e as limitações das técnicas de interpretabilidade utilizadas. Também será discutido como a combinação de modelos de inteligência artificial e módulos de interpretabilidade pode melhorar a capacidade preditiva e a compreensão dos modelos, facilitando assim a tomada de decisões orientada por dados no ambiente de negócios.

Processo de modelagem

O processo de modelagem é realizado em três fases: engenharia de dados, modelagem e análise de interpretabilidade do modelo.

1. Engenharia de dados

A engenharia de dados é a fase inicial em que o conjunto de dados é preparado e processado para ser usado na criação de modelos de inteligência artificial. Nesse caso, são executadas as seguintes ações:

- Definição do escopo da análise: neste caso, todos os funcionários que estiveram em licença médica nos últimos dois anos são considerados a população.
- Limpeza de dados: a qualidade dos dados é verificada e os registros com informações ausentes ou inconsistentes são removidos ou corrigidos.
- Transformação de variáveis: as variáveis categóricas são convertidas em variáveis numéricas usando técnicas como one-hot encoding ou ordinal encoding. Além disso, as variáveis numéricas são normalizadas ou padronizadas quando necessário.
- Seleção de variáveis: as variáveis mais relevantes para prever o desgaste dos funcionários são identificadas usando técnicas de seleção de variáveis, como correlação de Pearson, significância de recursos em modelos baseados em árvores ou eliminação recursiva de recursos.
- Construção de variáveis: novas variáveis são geradas a partir de variáveis existentes para analisar se elas são melhores preditoras do desgaste dos funcionários, como a "satisfação total", que foi construída como a soma das pontuações das variáveis "Satisfação com o ambiente", "Satisfação com o trabalho", "Avaliação de desempenho", "Equilíbrio entre vida pessoal e profissional", "Envolvimento no trabalho" e "Satisfação com as relações de trabalho".
- Divisão do conjunto de dados: o conjunto de dados é dividido em dois subconjuntos: treinamento e teste. O subconjunto de treinamento é usado para ajustar e otimizar os modelos de inteligência artificial, enquanto o subconjunto de teste é usado para avaliar o desempenho e a capacidade de previsão dos modelos.

2. Desenvolvimento do modelo

Nessa fase, diferentes modelos de inteligência artificial são treinados e validados usando o subconjunto de treinamento. Em particular, vários dos algoritmos de machine learning mais comuns, como regressão logística, árvores de decisão, máquinas de vetor de suporte, redes neurais e random forest, são ajustados e comparados para selecionar o modelo com o melhor desempenho.

Para evitar o treinamento excessivo e otimizar os hiperparâmetros dos modelos, são usadas técnicas de validação cruzada e pesquisa em grade ou aleatória. Além disso, foi dada atenção especial à complexidade do modelo durante o treinamento ao selecionar um algoritmo específico, a fim de facilitar sua interpretação.

Para isso, foi gerado um fluxo de desenvolvimento de modelo no ModelCraft™ (Fig. 13).

Para selecionar o modelo com a melhor capacidade de previsão, seu desempenho no subconjunto de teste é avaliado usando métricas como precisão, sensibilidade, especificidade e área sob a curva ROC (AUC-ROC). Essas métricas nos permitem avaliar a eficácia do modelo selecionado em termos de sua capacidade de prever corretamente o desgaste dos funcionários em dados não vistos anteriormente.

Considerando todos os aspectos, o algoritmo de random forest produz resultados de desempenho superiores, embora represente um desafio de interpretabilidade na compreensão de suas previsões. Esse modelo considerou 300 árvores de decisão e produziu uma precisão de 75% e uma sensibilidade de 84%. Portanto, essas são previsões muito confiáveis e os falsos negativos são raros. Isso é relevante para este estudo de caso, em que a empresa desejaria previsivelmente reduzir esse tipo de erro o máximo possível.

3. Análise de interpretabilidade

Nessa última fase, as técnicas de interpretabilidade são aplicadas para analisar e entender o comportamento e as decisões do modelo selecionado. Especificamente, os objetivos da análise foram:

— Entender quais variáveis são mais importantes na tomada de decisões da empresa em nível global, usando uma comparação por importância de cada variável.

— Entenda como as mudanças nas variáveis mais importantes afetam diferentes faixas populacionais.

— Entender os resultados do modelo em casos específicos em que uma certa probabilidade de desistência é observada.

Neste estudo de caso, as técnicas SHAP, LIME e PDP são usadas para explicar como o modelo toma decisões e como as entradas influenciam as previsões.

O SHAP permite obter resultados de interpretabilidade global, que dão uma interpretação da importância de cada variável, e o LIME permite realizar uma análise intuitiva da interpretabilidade local que permite explicar o resultado do modelo para cada funcionário com base em modelos lineares mais simples. Como complemento, os gráficos PDP permitem visualizar como as alterações em cada variável afetam a previsão do modelo.

Isso resultou na seguinte distribuição da importância de cada variável (Fig. 14).Nesse caso, observa-se que a variável com maior importância na previsão de desligamento (15,65%) é a "satisfação total", um indicador sintético definido como uma média ponderada de seis elementos (ambiente de trabalho, adequação das funções e áreas ao trabalho, classificação interna, conciliação familiar, relacionamento com colegas e supervisores, e cargo e responsabilidade do funcionário).

Esse resultado é intuitivo e mostra que a variável "satisfação total" foi bem projetada. No entanto, as três variáveis seguintes em termos de importância (tempo de serviço, salário e distância de casa ao trabalho) demonstraram ter grande influência na rotatividade de funcionários, que coletivamente é o dobro do indicador "satisfação total".

Para entender como cada variável é influenciada individualmente, os PDPs foram estudados (Fig. 15).

Em termos de tempo de serviço, a tendência se inverte após três anos: os funcionários com tempo de serviço intermediário são, em média, os menos propensos a deixar a empresa. Em relação à satisfação geral, observa-se uma tendência intuitiva: uma maior satisfação relatada em pesquisas internas resulta em uma menor taxa de desligamento.

Para complementar a análise anterior, o LIME foi usado para analisar, caso a caso, os valores das variáveis que influenciam a probabilidade de saída de determinados funcionários. A Fig. 16 mostra dois funcionários com diferentes probabilidades de saída obtidas com o modelo. O LIME mostra uma métrica de

Fig. 14. Interpretabilidade geral do modelo random forest usando SHAP, em que os valores de Shapley são usados para obter a importância das variáveis.

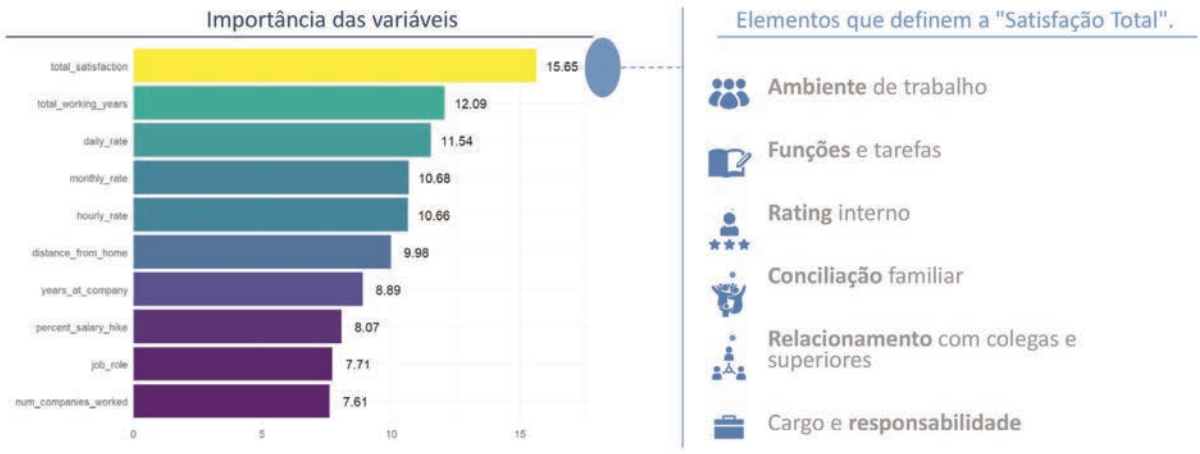
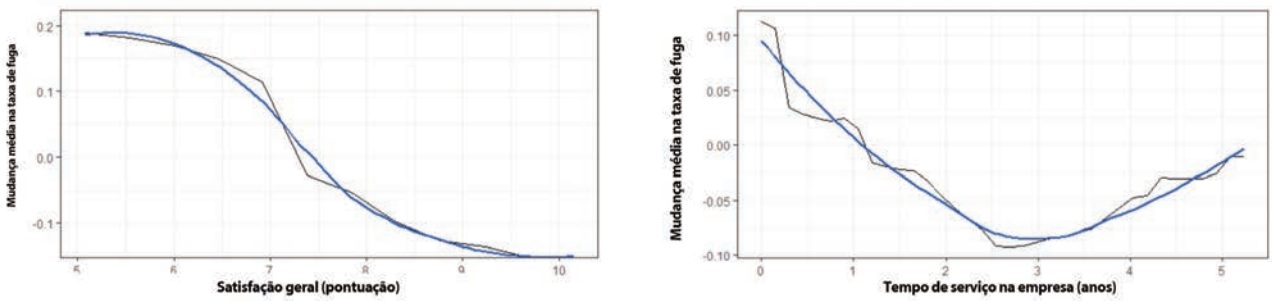


Figura 15. Gráficos PDP para as variáveis "satisfação total" e "tempo de serviço".



explicabilidade que representa a qualidade do ajuste linear obtido usando o modelo substituto local para explicar essas previsões.

Vale ressaltar que as causas mais relevantes de desligamento nesses dois casos não correspondem necessariamente às variáveis mais influentes em nível global. Embora se possa observar que a satisfação total contribui para explicar a probabilidade de saída do funcionário no caso 1, ela não parece ter um impacto significativo no caso 2, em que a probabilidade de saída é maior.

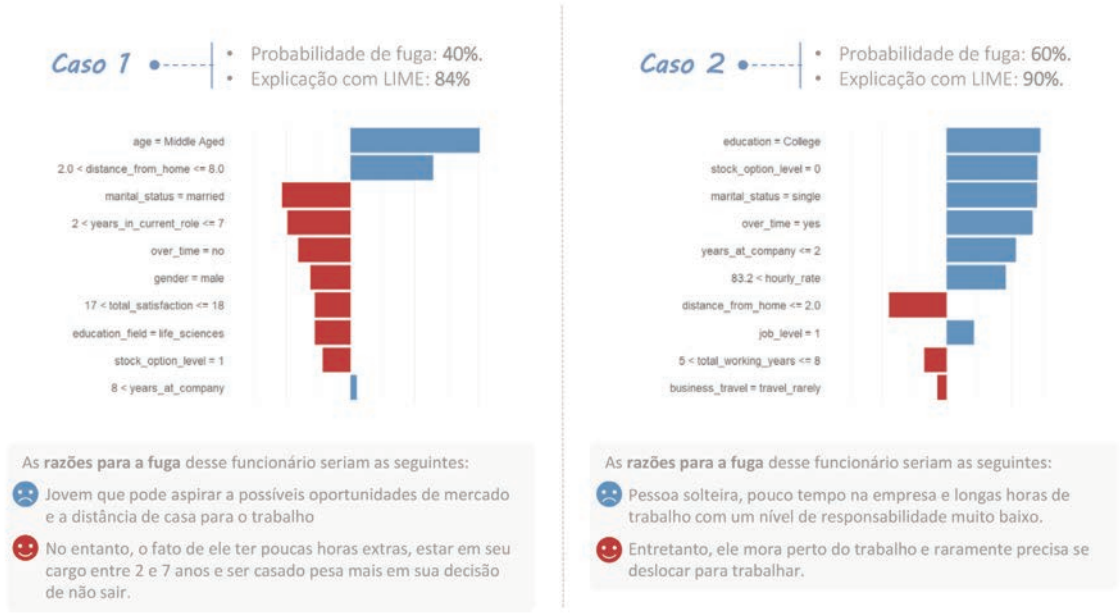
Isso reflete as dificuldades de interpretação desse modelo, que pode ser generalizado para modelos semelhantes: embora a satisfação total possa explicar a probabilidade de desistência em média, essa conclusão é uma generalização; há casos individuais e de grupo em que a desistência é explicada em maior grau por outras variáveis.

Conclusões do estudo de caso

Várias conclusões e lições aprendidas podem ser extraídas do estudo de caso de interpretabilidade de inteligência artificial apresentado, o que pode ser útil em aplicações futuras de modelos de AI e XAI:

- **Aplicação do modelo:** a aplicação e a interpretação corretas do modelo, nesse caso, podem ajudar a prever e evitar o desgaste dos funcionários. Entre os usos que podem ser feitos do modelo está a capacidade de criar diferentes perfis com propensão a sair e de identificar as características desses funcionários com antecedência para

Figura 16. Interpretabilidade local do modelo de random forest usando LIME.



tomar as medidas adequadas, o que, a longo prazo, pode contribuir para reduzir o nível de rotatividade na empresa.

- **Escolha do modelo:** o processo de modelagem mostrou a importância de comparar e validar diferentes algoritmos de machine learning para selecionar o modelo com a melhor capacidade de previsão. Nesse caso, o modelo random forest provou ser o mais adequado para prever a fuga dos funcionários.

- **Importância da interpretabilidade:** a aplicação de técnicas de interpretabilidade, como SHAP, LIME e PDP, proporcionou uma compreensão mais profunda de como o modelo toma decisões e como as entradas influenciam as previsões. Essas informações são cruciais para validar a aplicabilidade do modelo no contexto real e para garantir que as previsões sejam baseadas em recursos relevantes e significativos.

- **Variáveis de influência:** a análise de interpretabilidade identificou as variáveis mais relevantes para prever o desgaste dos funcionários. Essas variáveis podem ser úteis no desenvolvimento de estratégias de retenção e na melhoria da satisfação no trabalho. Além disso, a compreensão de como essas variáveis interagem entre si e como afetam diferentes segmentos da população de funcionários pode enriquecer a análise e facilitar a tomada de decisões com base em dados.

- **Implementação prática:** o estudo de caso demonstra a viabilidade e a utilidade da aplicação de técnicas de AI e XAI em um cenário realista, usando dados fictícios, mas representativos de uma situação comercial. Essa abordagem pode ser adaptada a outros contextos e problemas comerciais, aproveitando a inteligência artificial e a interpretabilidade para melhorar a tomada de decisões e obter resultados mais eficientes e eficazes.

- **Limitações:** ao mesmo tempo, esse caso de uso destacou as limitações e dificuldades na aplicação de técnicas de interpretabilidade post-hoc. É importante reconhecer que os métodos de interpretabilidade não são infalíveis e, às vezes, podem apresentar resultados aproximados ou parciais. Portanto, é essencial aplicar uma abordagem crítica e rigorosa ao interpretar e validar os resultados das técnicas de interpretabilidade.

- **Combinação de modelos de AI e módulos de interpretabilidade:** este estudo de caso mostra como a integração de modelos de AI e módulos de interpretabilidade pode melhorar a capacidade de previsão e a compreensão dos modelos. Isso facilita a adoção de soluções baseadas em AI na tomada de decisões de negócios.

- **Continuidade na análise de interpretabilidade:** por fim, deve-se enfatizar que a análise de interpretabilidade não deve ser um exercício único aplicado durante o desenvolvimento do modelo, mas deve ser realizada de forma contínua, reproduzível e confiável durante toda a vida útil do modelo.

Concluindo, este estudo de caso de interpretabilidade em inteligência artificial proporcionou uma experiência valiosa na aplicação de técnicas de AI e XAI em um contexto comercial e mostrou o potencial da AI e da interpretabilidade para aprimorar a tomada de decisões, revelando, ao mesmo tempo, as limitações e dificuldades associadas a essas técnicas e a necessidade de uma abordagem crítica e rigorosa ao interpretar e validar os resultados da AI.

Conclusão

Com a programação correta, um computador pode se tornar um teatro, um instrumento musical, um livro de referência, um oponente de xadrez. Nenhuma outra entidade no mundo, a não ser o ser humano, tem uma natureza tão adaptável e universal.

Daniel Hillis[74]

Este estudo apresentou a Inteligência Artificial Explicável (XAI), seus fundamentos, contexto e técnicas para melhorar a interpretabilidade dos modelos. Foram discutidos os principais desafios enfrentados pelos modelos de inteligência artificial em termos de interpretabilidade e como a tecnologia pode ajudar a resolvê-los, incluindo um estudo de caso desenvolvido com o ModelCraft™ para demonstrar como essas técnicas podem ser empregadas para entender e explicar os modelos de AI.

A disciplina de AI, e dentro dela a XAI, cresceu em importância globalmente nos últimos anos, já que o desenvolvimento de tecnologias de AI de alto desempenho se tornou uma prioridade para muitos setores, da saúde à segurança, dos serviços financeiros à energia, entre outros. A interpretabilidade surge como uma necessidade para entender e aprimorar os modelos de AI, o que é particularmente complexo para determinadas técnicas.

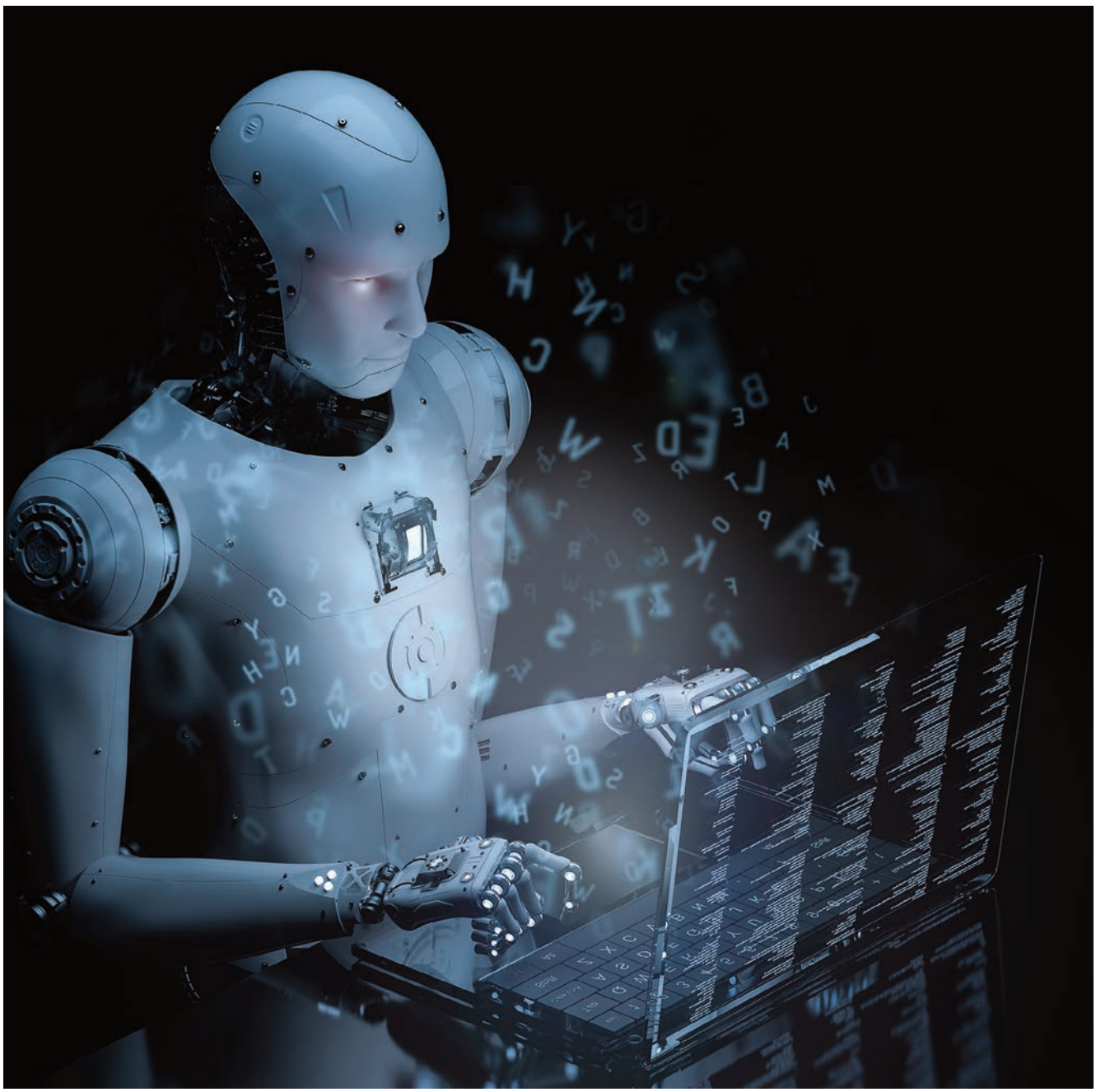
Como vimos, os modelos de AI podem enfrentar dificuldades para explicar os resultados ou a lógica por trás de suas decisões. Isso ocorre porque esses modelos usam técnicas de aprendizagem profunda e algoritmos complexos para aprender com os dados, que muitas vezes são difíceis de interpretar, o que apresenta desafios na avaliação dos modelos de AI e na confiabilidade de seus resultados.

Como resultado, o framework regulatório de AI está evoluindo rapidamente, e espera-se que as organizações se adaptem aos novos requisitos de transparência, explicabilidade e equidade no uso de modelos de AI. Isso implica a necessidade de uma abordagem holística que integre a interpretabilidade e a explicabilidade na organização e nos processos de cada empresa, abrangendo técnicas de interpretabilidade, gestão de riscos de modelos, colaboração interdisciplinar e treinamento em XAI para profissionais envolvidos no desenvolvimento e na aplicação da AI, entre outros.

Concluindo, a interpretabilidade dos modelos de inteligência artificial é uma área emergente de pesquisa e espera-se que continue a se desenvolver e a crescer em importância à medida que os modelos de AI se tornam mais complexos, a regulamentação continua a proliferar e seu uso se estende a domínios mais altamente sensíveis.

[74]Daniel Hillis (n. 1956), inventor, empresário e cientista americano, pioneiro da computação paralela e de seu uso no campo da inteligência artificial, com mais de 300 patentes publicadas.

Glossário



Aprendizado de máquina (*machine learning*): subcampo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem que as máquinas aprendam e melhorem seu desempenho em uma tarefa específica por meio da experiência.

Caixa branca (*white box*): sistema ou modelo de AI cujo funcionamento interno é simples de entender e explicar.

Caixa preta (*black box*): sistema ou modelo de AI cujo funcionamento interno é desconhecido ou difícil de entender.

Direito a uma explicação: conceito jurídico que sustenta que os indivíduos têm o direito de saber como são tomadas as decisões automatizadas que os afetam e de receber uma explicação compreensível de como funcionam os algoritmos envolvidos.

Explicabilidade: capacidade de um sistema de AI de fornecer justificativas claras e compreensíveis de suas previsões ou decisões aos usuários e às partes interessadas. Isso envolve o fornecimento de informações detalhadas e contextualizadas sobre como e por que um modelo de AI chega a uma determinada conclusão, facilitando a confiança e a adoção da tecnologia.

GPT-4: quarta geração do modelo Generative Pre-trained Transformer, desenvolvido pela OpenAI Foundation, que é usado para tarefas de processamento de linguagem natural e de geração de texto.

Inteligência artificial (AI): campo de estudo que busca desenvolver sistemas capazes de realizar tarefas que normalmente exigem inteligência humana, como aprendizado, raciocínio, percepção e tomada de decisões.

Inteligência artificial explicável (XAI): abordagem de IA que busca tornar os modelos de inteligência artificial mais compreensíveis e transparentes para os seres humanos.

Interpretabilidade: facilidade com que os seres humanos podem entender o processo de tomada de decisão de um modelo de IA, bem como as relações entre os recursos de entrada e as previsões ou decisões. Um modelo interpretável permite que os usuários entender como se chega a uma previsão ou decisão específica.

LIME *(Local Interpretable Model-agnostic Explanations):* técnica de explicabilidade que ajuda a entender as previsões individuais de um modelo de AI criando aproximações locais interpretáveis.

Modelo sub-rogado: modelo interpretável que é treinado para imitar as previsões de um modelo de AI complexo e menos interpretável, como uma rede neural profunda. O objetivo de um modelo sub-rogado é fornecer uma explicação simplificada e compreensível de como o modelo original toma decisões.

Open AI Foundation: organização de pesquisa e desenvolvimento de inteligência artificial, atualmente participada da Microsoft, cujo objetivo declarado é garantir que a AI beneficie toda a humanidade.

Partial Dependence Plot (PDP): técnica de visualização que mostra o efeito médio de uma característica nas previsões de um modelo de AI, mantendo constantes todas as demais características. Ela ajuda a entender a relação entre características e previsões e a detectar possíveis interações e não linearidades.

Rede neural profunda: tipo de algoritmo de machine learning que consiste em várias camadas de neurônios artificiais e é capaz de aprender representações hierárquicas de dados.

Regulamento Geral de Proteção de Dados (GDPR): legislação da União Europeia que estabelece regras para a coleta, o armazenamento e o processamento de dados pessoais de cidadãos da UE.

SHAP (***SHapley Additive exPlanations):*** técnica de explicabilidade que usa valores de Shapley da teoria dos jogos cooperativos para atribuir a importância de cada variável na previsão de um modelo de AI.

Sparsity: propriedade de um modelo em que ele considera apenas o subconjunto de variáveis que são realmente relevantes para a estimativa.

Teste de esquema de Winograd: teste de compreensão de linguagem natural que avalia a capacidade de uma AI de resolver ambiguidades na linguagem por meio do uso de conhecimento e raciocínio comuns.

Teste de Turing: teste proposto por Alan Turing em 1950 que avalia a capacidade de uma máquina de imitar a inteligência humana a ponto de ser indistinguível de um ser humano em uma conversa.

Transformer: arquitetura de rede neural introduzida pelo Google Brain em 2017, usada principalmente em tarefas de processamento de linguagem natural (NLP). Os transformers são conhecidos por sua capacidade de lidar com longas sequências de dados e por sua eficiência no treinamento. Eles se baseiam em mecanismos de atenção, que permitem que a rede pondere a importância relativa de palavras ou elementos em uma sequência ao longo do tempo. Os transformers impulsionaram o desenvolvimento de modelos de linguagem de última geração, como o GPT e o BERT, e revolucionaram o campo da NLP.

Transparência: abertura e acessibilidade de um sistema de AI em termos de seu desenho, estrutura e processos internos. Um sistema transparente permite que os usuários e as partes interessadas examinem e entendam seus componentes, algoritmos e decisões.

Referências

Broniatowski, D. (2021). Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. https://www.nist.gov/publications/psychological-foundations-explainability-and-interpretability-artificial-intelligence

Comisión Europea (2021). Artificial Intelligence Act / Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión. https://artificialintelligenceact.eu/

Comisión Europea (2019). Dirección General de Redes de Comunicación, Contenido y Tecnologías, Directrices éticas para una IA fiable, Oficina de Publicaciones, 2019, https://data.europa.eu/doi/10.2759/14078

C. Rudin, C. Chen, Zhi Chen, H. Huang, L. Semenova, C. Zhong. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. http://essay.utwente.nl/91965/

Doshi-Velez, F., et al. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608

Devis (2011). https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.html

Dimensions (2022). https://app.dimensions.ai/discover/publication

EBA (2021). Discussion paper on machine learning for IRB models. https://www.eba.europa.eu/regulation-and-policy/model-validation/discussion-paper-machine-learning-irb-models

European Parliamentary Research Service (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530

Floridi et al. (2022). capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." Annals of statistics (2001): 1189-1232. https://www.jstor.org/stable/2699986

Gall, R. (2018). Machine Learning explainability vs interpretability: two concepts that could restore trust in AI, KDnuggets. https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

GDPR (2018), Recital 71. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Goldstein, A.; Kapelner, A.; Bleich, J; Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. https://arxiv.org/abs/1309.6392

Harnad, D. (2003). Can a machine be conscious? How? https://web-archive.southampton.ac.uk/cogprints.org/5330/

IBM (2022). Explainable AI (XAI). https://www.ibm.com/watson/explainable-ai

iDanae (2022). ML Applied to Credit Risk: building explainable models. Quarterly Newsletter 3Q22. iDanae Chair. https://blogs.upm.es/catedra-idanae/wp-content/uploads/sites/698/2022/10/Idanae-3Q22.pdf

Jonathon Phillips, P.; Hahn, H.; Fontana, P; Yates, A.; Greene, K. K.; Broniatowski, D. A.; Przybocki, M. A. (2021). Four Principles of Explainable Artificial Intelligence. NIST. https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence

Kaggle (2017). IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

LeCun, Y.; Bengio, Y.; Hinton, G. (2015). Deep learning. Nature. https://pubmed.ncbi.nlm.nih.gov/26017442/

Leventi-Peetz, A.-M., et al. (2022). Deep Learning Reproducibility and Explainable AI (XAI). https://arxiv.org/abs/2202.11452

Levesque, H. (2014). On our best behaviour. Written version of the Research Excellence Lecture presented in Beijing at the IJCAI-13 conference. Artificial Intelligence, vol. 212, pages 27-35. https://doi.org/10.1016/j.artint.2014.03.007

Lundberg, S. M.; Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. https://dl.acm.org/doi/10.5555/3295222.3295230

Management Solutions (2023). ModelCraft. Modelización por componentes. https://www.managementsolutions.com/es/microsites/soluciones-propietarias/modelcraft

Management Solutions (2022). Gamma. Sistema de gobierno de modelos. https://www.managementsolutions.com/es/microsites/soluciones-propietarias/gamma

Management Solutions (2021). Nota técnica sobre el EBA Discussion paper on machine learning for IRB models. https://www.managementsolutions.com/es/publicaciones-y-eventos/apuntes-normativos/notas-tecnicas-normativas/documento-de-debate-sobre-machine-learning-en-el-enfoque-irb

Management Solutions (2020). Auto machine learning, towards model automation. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/auto-machine-learning-towards-model-automation

Management Solutions (2018). Machine learning, a key component in business model transformation. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/machine-learning-a-key-component-in-business-model-transformation

Management Solutions (2015). Data science and the transformation of the financial industry. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/data-science

Marcinkevics, R. (2020). Interpretability and Explainability: A Machine Learning Zoo Mini-tour. ETH Zürich, Department of Computer Science, Institute for Machine Learning. https://arxiv.org/abs/2012.01805

McCarthy, J. (2004). What is artificial intelligence? Stanford University. http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell.2019,267, 1–38. https://www.sciencedirect.com/science/article/pii/S0004370218305988

OECD (2019). Principles for Artificial Intelligence. https://www.oecd.org/digital/artificial-intelligence/

Oneto, L., Chiappa, S., (2020). Fairness in Machine Learning. 2012.15816.pdf (arxiv.org)

Ribeiro, M. T.; Singh, S.; Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. https://arxiv.org/abs/1602.04938

Ribeiro, M. T.; Singh, S.; Guestrin, C. (2018). "Anchors: high-precision model-agnostic explanations". AAAI Conference on Artificial Intelligence (AAAI). https://ojs.aaai.org/index.php/AAAI/article/view/11491

Roscher, R.; Bohn, B.; Duarte, M.; Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. https://ieeexplore.ieee.org/document/9007737

Shapley, L. (1953). A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307-317. https://doi.org/10.1515/9781400881970-018

Sudjianto, A.; Knauth, W.; Singh, R.; Yang, Z.; Zhang, A. (2011). Unwrapping The Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification. Cornell University. https://arxiv.org/abs/2011.04041

Sudjianto, A.; Zhang, A. (2021). Designing Inherently Interpretable Machine Learning Models. https://arxiv.org/abs/2111.01743

Turing, A. (1950). Computing Machinery and Intelligence. Mind 49: 433-460.

Vilone G., Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, vol. 76: 89-106. https://www.sciencedirect.com/science/article/pii/S1566253521001093

White House OSTP (2022). Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Yang, Z.; Zhang, A.; Sudjianto, A. (2019). Enhancing Explainability of Neural Networks through Architecture Constraints. https://arxiv.org/abs/1901.03838

Zhou, N.; Zhang, Z.; Nair, V. N.; Singhal, H.; Chen, J.; Sudjianto; A. (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. https://arxiv.org/abs/2105.06558

Observação: esta publicação foi produzida com a ajuda de várias ferramentas de inteligência artificial (AI). Essas ferramentas foram usadas para várias tarefas, como busca de informações, coleta e organização de dados e geração de resumos. De qualquer forma, o material final desta publicação foi escrito por uma pessoa e não por uma AI.

Nosso objetivo é superar as expectativas dos nossos clientes sendo parceiros de confiança

A Management Solutions é uma empresa internacional de serviços de consultoria com foco em assessoria de negócios, riscos, organização e processos, tanto sobre seus componentes funcionais como na implementação de tecnologias relacionadas.

Com uma equipe multidisciplinar (funcionais, matemáticos, técnicos, etc.) de mais de 3.300 profissionais, a Management Solutions desenvolve suas atividades em 44 escritórios (19 na Europa, 21 nas Américas, 2 na Ásia,1 na África e 1 na Oceania).

Para atender às necessidades de seus clientes, a Management Solutions estruturou suas práticas por setores (Instituições Financeiras, Energia e Telecomunicações) e por linha de negócio, reunindo uma ampla gama de competências de Estratégia, Gestão Comercial e Marketing, Gestão e Controle de Riscos, Informação Gerencial e Financeira, Transformação: Organização e Processos, e Novas Tecnologias.

A área de P&D presta serviço aos profissionais da Management Solutions e a seus clientes em aspectos quantitativos necessários para realizar os projetos com rigor e excelência, através da aplicação das melhores práticas e da prospecção contínua das últimas tendências em *data science*, *machine learning*, modelagem e *big data*.

Javier Calvo Martín
Sócio
javier.calvo.martin@managementsolutions.com

Manuel Ángel Guzmán Caba
Sócio
manuel.guzman@managementsolutions.com

Segismundo Jiménez Láinez
Gerente
segismundo.jimenez@msspain.com

Luz Ferrero Peña
Supervisora
luz.ferrero@msgermany.com.de

Management Solutions, serviços profissionais de consultoria

Management Solutions s é uma firma internacional de serviços de consultoria focada na assessoria de negócio, riscos, finanças, organização e processos

Para mais informações acesse **www.managementsolutions.com**

Siga-nos em:

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña Málaga London Frankfurt Düsseldorf Paris Amsterdam Copenhagen Oslo Warszawa Wroclaw Zürich Milano Roma Bologna Lisboa Beijing Istanbul Johannesburgo Sydney Toronto New York New Jersey Boston Pittsburgh Atlanta Birmingham Houston San Juan de Puerto Rico San José Ciudad de México Monterrey Querétaro Medellín Bogotá Quito São Paulo Río de Janeiro Lima Santiago de Chile Buenos Aires